EARLY DETECTION OF MORBIDITY IN FEEDLOT CATTLE USING PATTERN RECOGNITION TECHNIQUES

A Thesis Submitted to the College of Graduate Studies and Research in Partial Fulfillment of the Requirements for the Degree of Master of Science in the Department of Agricultural Bioresource Engineering

University of Saskatchewan

Saskatoon, SK

By

Réka Silasi

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my theses.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Agricultural Bioresource Engineering

University of Saskatchewan

Saskatoon, Saskatchewan S7N 0W0

ABSTRACT

Computer algorithms are routinely used to aid in the identification of biological patterns not easily detected with standard statistics. Currently, observed changes in normal patterns of feeding behavior (FB) are used to identify morbid feedlot cattle. The objective of this study was to use pattern classification techniques to develop algorithms capable of identifying morbid (M) cattle earlier than traditional pen checking methods. In two separate studies, individual feeding behaviour was obtained from 384 feedlot steers (228± 22.7 kg, initial BW) in a 226 d trial (model dataset), and 384 feedlot heifers (322 ± 34.7 kg, initial BW) in a 142 d trial (naive dataset). Data was collected using an automated feed bunk monitoring system. FB variables calculated included feeding duration, inter-meal interval (min., max., avg., SD and total; min/d) and feeding frequency (visits/d). Animal health records including the number of times treated, d in the hospital and d on feed were also collected. Ninety-three and 53 morbid (M) animals were identified in each trial respectively, and were categorized into low, moderate and high groups, based on severity of sickness. FB data for 68 cattle from the model dataset (45 classified as Moderate and 25 classified as High) was analyzed to develop an algorithm which would aid in identifying morbid FB. This algorithm was later tested on 18 M animals (12 classified as Moderate and 6 as High) in the naive dataset. The pattern recognition procedure involved reducing data dimensionality via Principal Component Analysis, followed by K-means clustering and finally the development of a binary string to aid in the classification of M feeding behaviour. The developed procedure resulted in an overall classification accuracy of 84 % (82.5 and 85 % accuracy for H and M, respectively) for the model dataset, and 75 % overall (100 and 50 % accuracy for H and

M, respectively) for the naive dataset. The model predicted morbidity on average 3.3 and 1.2 d earlier than pen checkers could for each trial respectively. The application of pattern recognition algorithms to FB shows value as a method of identifying morbid cattle in advance of overt physical signs of morbidity.

ACKNOWLEDGEMENTS

I am particularly grateful to my co-supervisor, Dr. Karen Schwartzkopf-Genswein for her guidance, helpful and constructive comments and patience throughout this journey. I would like to thank the rest of my committee members, Drs. Trever Crowe, Tim McAllister and Dr. Ron Bolton for their support and assistance. Also, I would like to express my gratitude to Mr. Bernie Genswein for his expert advice and help. I wish to thank my friends for their encouragement, and most importantly, this work would have not been possible without the moral support of my family.

Mamusnak, Papusnak,Gergőnek

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABREVIATIONS

AI	Artificial intelligence
Avg.	Average
BRD	Bovine respiratory disease
BW	Body weight
CLS	Confidence Level of Sickness
CUSUM	Cumulative Sums Analysis
D	Day
DM	Dry Matter
FB	Feeding behaviour
H	Healthy
Hi	High
IRAD	Integrated Research Analysis Database
KDD	Knowledge Discovery in Databases
Lo	Low
M	Morbid
Max.	Maximum,
Mo	Moderate
Min.	Minimum
PC	Principal component
PCA	Principal Component Analysis
RFID	Radio frequency identification
SD	Standard deviation
U	Unknown

1. INTRODUCTION

"Even the recognition of an individual whom we see every day is only possible as the result of an abstract idea of him formed by generalization from his appearances in the past."

Games G. Frazer

The word "recognition" plays a significant role in our daily lives as it is a basic procedure practiced by all human beings. Many professions, businesses and enterprises depend on individuals or machines to correctly recognize and identify pre-defined objects, living organisms or behaviours. Comparing an object or situation against existing knowledge stored in the human mind is a complex and multi-dimensional task and involves information gathering and precise comparisons on various levels. For example, when we see a cow, we first recognize that it is an animal. Then we look at specifics such as its size, color, shape and position of its head in relation to its body, and so on. We may have seen many cows before, and learned what they 'should' look like. After assessing its attributes, we make comparisons of this animal with the existing images stored in our mind leading to the conclusion that it is indeed a cow. A pen checker in a feedlot is expected not only to recognize the type of animal correctly, but also its state of health. Recognizing the health status of cattle can be difficult, as it is a subjective procedure based on behavioural rather than physical characteristics (Broom, 2006). Primary among these behavioural characteristics is feeding behaviour. One of the first indicators that an animal is sick is that it is 'off-feed' (Edwards, 1980). It is known that feeding behaviour of cattle is affected by various factors such as feed availability,

weather, social interactions and the health status of individuals. Appetite depression is one of the most important early symptoms associated with feedlot diseases and disorders (Blezinger, 2005; Hutcheson, 1988). A reliable method of recognizing patterns of feeding behaviour typical of cattle morbidity or proneness to disease would be of tremendous value to the feedlot industry because of the direct relationships between animal health and welfare, feed intake, and economic return. Establishment of such a knowledge base would enable pen checkers to assess observed behaviours relative to reliable reference standards and thereby improve the accuracy of identification of sick animals.

This thesis discusses the use of pattern recognition techniques on cattle feeding behaviour, and introduces a proposed automated method to identify cattle morbidity in its early stages, before the physical characteristics of sickness become evident. Despite the potential benefit of this strategy, few attempts have been made to develop automatic or semi-automatic tools for post-processing of feeding behaviour data. Automation refers to a computerized system programmed to recognize feeding behaviour patterns developed from feed intake and health management data that are associated with existing or developing morbidity among feedlot cattle.

The need to process feeding behaviour data automatically became evident after the introduction into commercial settings of an automated behaviour monitoring system (GrowSafe Systems ™, Airdrie, AB) that is based on radio frequency identification (RFID). At Cactus Feeders (Amarillo, TX), this system has generated datasets believed

to be the largest and most complete datasets on feeding behaviour of sick and healthy cattle in the world.

The collection system captures ‘true’ feedlot behaviour by cattle, free of artifacts introduced by human or technical intervention. The proposed model combines data analysis, i.e., understanding of behavioural data processing and signal recognition, with common pattern recognition techniques to provide insightful biologically meaningful solutions. By understanding and replicating the manner in which humans interpret feeding behaviour, the ultimate goal is to use perceptual computer models to classify feeding behaviour as healthy or morbid.

2. LITERATURE REVIEW

Chapter 2 will include a review of the current management techniques implemented in the intensive beef production industry. Special attention will be given to animal health and wellfare issues, and to the method used for the early detection of sickness, as these approaches are still in their formative stages. The work is based on theories of animal behaviour, along with the employment of pattern recognition through artificial intelligence (AI). The development of a novel process to allow computers to process and analyze feeding behaviour data in a manner similar to that which is performed by experienced feedlot personnel is described. The process consists of several programs that apply AI techniques to feeding behaviour measurements that are in turn used to understand the behaviour and health state of individual feedlot cattle. The following sections highlight the motivation behind, and the objectives of the research depicted in this thesis.

2.1. The Feedlot Industry

The beef industry is a large contributing factor to the world in terms of economy, nutrition and the environment. Cattle have been consumed around the world for centuries, and today beef production, consumption, imports and exports continue to follow their recent trend of annual historic heights (FAS, 2007).

In North America, beef cattle are born on cow-calf production farms. Cow-calf production is the first stage of the beef production cycle, and it is the most traditional phase in the cattle-beef commodity chain (MacLachlan, 2001). At this point, calves are

raised on pasture together with their mother and are weaned at 6 to 9 months of age, between 225 and 325 kg (Mathison, 1993) and transported to a feedlot.

A feedlot is an area designated for housing and fattening cattle for the market. Generally, feedlots in North America are comprised of multiple pens, a centrally located water system (within each pen), feed bunks and resting areas. Rows of pens are separated by alleys used for daily tasks such as feeding, pen cleaning and animal handling. The number of animals housed in each pen may be as high as 300+, but pen sizes vary widely among feedlots. The animals in each pen are typically homogeneous with respect to ownership, sex, breed, and size (MacLachlan, 2001). Feedlot capacity varies greatly, however economies of size are motivating the shift toward larger feed yards (Mintert, 2003). Some of the largest operations in North America have a one-time capacity of 25,000 head or more (MacLachlan, 2001), and achieve good economies of scale by reducing production cost per animal. On average, finishing cattle spend 120 days in the feedlot (MacLachlan, 2001), thus the turnover rate of such facilities is two to three times each year. Depending on breed, level of intake and diet composition, finished feedlot cattle range between 500 and 600 kg, gaining approximately 1-2 kg/d of body weight (Mathison, 1993). The health of feedlot cattle are heavily influenced by the experience of the feedlot management and staff.

2.1.1. Feedlot Management

Feedlot management has become a sophisticated, precise, and science-oriented task, and it is clear that feeding and management strategies have a decisive impact on cattle performance (Mintert, 2003). For example, of surveyed feedlot owners, feed bunk

management was considered to be a critical factor affecting feed intake and animal performance (Galyean, 1996). Several factors are believed to influence the nutritional needs of receiving cattle (Hutcheson, 1988; McEwen and Wingfield, 2003). Among these, the type of diet and feeding regime are of utmost importance to feedlot operators, as is the familiarization of the animals to their new surroundings, feed and feeding regime (Hutcheson *et al.*, 1997). In most feedlots, feed is delivered in a truck to a feed bunk (up to three times daily), which lines the front of each pen. The scheduled delivery of feed rations is important, as the availability of fresh feed at each feeding session assures that the cattle will eat and gain weight with optimal efficiency (Pritchard and Bruns, 2003).

One of the most challenging periods for feeding cattle is during the receiving period, a short period of time (30 – 40 d) following the arrival of cattle at the feedlot (Hutcheson and Cole, 1986). Receiving calves are fed a diet consisting mainly of forages, mixed with a small percentage of grain (70 and 30 %, respectively), with their rations gradually increasing in grain content (up to 90 %) (Mathison, 1993; Muir *et al.*, 1998). The combination of feedstuffs used in a finishing ration often changes due to several factors such as relative price, animal breed and the experience of the feedlot staff (Mathison, 1993).

2.1.2. Receiving Calves

The receiving of new cattle into the facility requires careful planning and management, as newly arrived calves are often tired, hungry and thirsty. Upon entry to the feedlot, the animals undergo management procedures which may include hot-iron

branding, castration, dehorning, vaccination and treatment for internal and external parasites (Radostits, 1996). Animals may also be mass medicated with pharmaceuticals such as antibiotics. For example, in the United States more than 90 % of feedlot operators administer vaccines and antibiotics upon arrival of young cattle at the feedlot (NAHMS, 1999). Performance enhancing hormone implants are also administered to increase average daily gain and improve feed efficiency (Roeber *et al.*, 2000). It is believed that as a result of such extensive handling procedures, the animals' homeostasis may be challenged and could be disturbed, resulting in stress and an increased susceptibility to disease (McEwen and Wingfield, 2003).

2.1.2.1. Stress

Homeostasis is a term that refers to to 'being in balance.' The inability to maintain homeostatic balance results in the development of stress (Sapolsky, 2000). Stress is defined as a non-specific response of the body to any demand from the environment (Selye, 1955). It is well documented that the physiological response to stressors varies greatly among animals, and it has been argued that this variability can be accounted for by differences in vulnerability to stressors. In his 2005 review paper, Sapolsky mentioned two types of stressors: physical and psychosocial. He defined a physical stressor as an external challenge to homeostasis, whereas a psychosocial stressor as the anticipation (justified or not) that a challenge to homeostasis looms. Receiving calves are exposed to both physical and psychosocial stressors upon arrival to the feedlot (Cole and Hutcheson, 1990; Hodgson *et al.*, 2005; Hutcheson, 1988; Johnson, 1985).

2.1.2.2. Stressors

Separating the calf from its mother is assumed to impose a great amount of stress on the calf and dam alike (Loerch and Fluharty, 1999). "Preconditioning" is a term used in the feedlot industry and refers to a calf management program geared towards reducing disease incidence, with the goal of improving the growth performance of freshly weaned calves (Pate and Crockett, 2002). Although preconditioning has been suggested to decrease weaning stress (Pate and Crockett, 2002), it is seldom implemented because of cost and/or lack of adequate facilities (Macartney, 2003). Additional stressors that young calves are exposed to include marketing through auction barns, transportation, exposure to new environments, commingling with other animals, handling, and consumption of novel feed (Galyean *et al.*, 1981; Galyean and Hubbert, 1995; Grandin, 1997; Hutcheson, 1988; Loerch and Fluharty, 1999). To gain a better appreciation of the stressors that animals routinely face, one only needs to consider the transport of cattle to the feedlot. Although the length of the trip may vary with location, in most cases the trip is divided into two transportation events, the first half being from the cow-calf producer to the auction market, followed by a second trip to the feedlot. Typically, at the auction market the calves are unloaded from the transport vehicle and commingled with animals from other sources. The mixing of cattle from different sources may expose animals to a variety of infectious agents. The feedlot environment after arrival may also impose additional stressors as the animals often have to acclimate to mud, manure, exposure to a new social environment and novel feed (Herskin *et al.*, 2003; Loerch and Fluharty, 1999). Sometimes animals are not fed for several days before reaching their final destination, and research has shown that despite the animals being hungry, feed intake

of newly arrived cattle is usually low (Cole, 1982; Cole, 1996; Hutcheson, 1988). Furthermore, the establishment of social and dominance order within each pen may also inflict problems, as animals of distinct ranks experience different patterns and levels of stress (Sapolsky, 2005). These findings provide strong evidence that stressors have a direct effect on feeding behaviour and performance, and consequently on herd health and efficiency (Cole, 1982; Loerch and Fluharty, 1999). Chronic stress can cause immunosupression, leaving the animal vulnerable to infectious agents (McNamara and Buchanan, 2005; Sapolsky, 2005).

2.1.3. Animal Health

Infectious diseases are a significant concern to the livestock industry in terms of animal welfare and feedlot economy (Duff and Galyean, 2007; Gardner *et al.*, 1999). Therefore the control of such diseases must be considered in any herd health management program. Feedlots deal with health-related concerns on a daily basis, and it is well known that newly arrived calves account for the majority of disease control and management issues (Duff and Galyean, 2007). The morbidity rate is generally much higher for calves (30 to 50 %) than for older animals (less than 30 %), (Johnson, 1985) as low feed intakes may compromise the animal's immune system leading to poor health and growth performance (Cole, 1982; Forbes, 2003; Rivera *et al.*, 2005). Subtheraputic use of antimicrobials and other various feed additives may offset some negative impacts of stress on health and growth (Hardy, 2002). In support of such practice, Phillips *et al.* (2004) and Rivera *et al.* (2005) suggest that continued use of antibiotics as a feed additive reduces mortality and morbidity rates at the feedlot, and

improve growth and feed efficiency. However, others argue that the widespread use of antibiotics can cause the development of antibiotic-resistant bacteria, which not only affects the animal but may also have implications for human health (Kumar *et al.*, 2005). For a more targeted discriminatory use of antibiotics, the development of a technique for the early identification of animals that are infected and/or prone to disease would be valuable (Blezinger, 2005). However, the exact diagnosis of subclinical infection is a major problem, as current methods rely on identification based on physical symptoms shown by the animal (Galyean *et al.*, 1999; Gardner *et al.*, 1999). The development of a technique that would be able to easily and promptly identify morbid animals would likely increase the efficacy of antibiotics, as they could be implemented earlier in the disease cycle. Treatment records indicate that the earlier a sick animal is identified, the better its chances of survival (Smith, 2005).

2.1.3.1. Bovine Respiratory Disease

Bovine respiratory disease (BRD) is one of the most prominent feedlot health issues (Duff and Galyean, 2007; Loneragan, 2001; Pinchak *et al.*, 2004; Smith, 1998). It is a disease of the respiratory tract, caused by stress, viral and bacterial infections, and numerous other stressors and agents such as dust, cold and fatigue (Bagley, 1997; Griffin, 1998; Loerch and Fluharty, 1999). BRD is of significant concern to feedlot operators in terms of animal welfare and economic loss (Duff and Galyean, 2007; Loneragan, 2001; Macartney, 2003) This condition accounts for 65-77 % of morbidity and 44-72 % of mortality rates in the United States (Edwards, 1996; Galyean *et al.*, 1999; Quimby, 2001). Approximately 65 to 80 % of BRD occurs during the first 45 days

in the feedlot (Griffin, 1998; Loneragan, 2001; Mathison, 1993; Smith, 1998). Physical signs of an animal having BRD include thick nasal discharge, difficulty breathing, discharge from eyes, red peeling muzzle and listless behaviour (Galyean *et al.*, 1999; Griffin, 1998). Body temperature of individuals with BRD is also frequently elevated to 39.4°C or above (Griffin, 2006). The normal range of cattle body temperature varies due to various factors such as the animal's environment, time of day, and the activity level of the animal. The body temperature of healthy cattle can range within the margins of 37.8 - 39.4 °C (Encyclopaedia Britannica, 1965), with an average of 38.6° (Academic American Encyclopedia, 1994).

2.1.3.2. Pen checking

Identifying feedlot disease is not an easy task, as different diseases may cause different clinical symptoms and behavioural differences in individual cattle (Galyean *et al.*, 1999). Typically, trained pen riders scan each pen daily and visually inspect the animals. Animals that appear to be sick are taken to a hospital pen, where they are treated with antibiotics and monitored.

Identification of BRD is subjective and not always accurate. Despite taking daily measures, clinical signs of disease often still go undetected (Gardner *et al.*, 1999). For example, abattoir records show that 68 % of untreated animals had lung lesions at slaughter (a sign that the animal had respiratory disease at some point during its life) (Gardner *et al.*, 1999; Wittum *et al.*, 1996). Diseases caused by bacterial infections are often treated by the administration of antibiotics, but these drugs are ineffective for viral pathogens. In addition, the economic losses associated with the disease do not stop with

the cost of antibiotic treatment as extra labour is required to deal with diseased animals and growth performance and carcass quality are also frequently compromised (Galyean, 1999; Larson, 2005; Loneragan, 2001; Rivera *et al.*, 2005; Smith, 1998). Feedlots may be able to reduce health problems by planning a more sophisticated and unbiased health maintenance and disease prevention program, as visual surveillance alone is unlikely to be the best method of early detection of morbidity.

2.1.3.3. Detection of BRD

One of the key behaviours pen checkers assess to identify sick animals is feeding behaviour (Edwards, 1980). Generally, pen checkers make their rounds around feed delivery. Prior to feed delivery, Pavlov's principle seems to occur as cattle anticipate feed delivery (Sowell *et al.*, 1999). Anecdotal evidence suggests that healthy animals stay true to this phenomenon; whereas sick animals don't await feed delivery and often do not react to the arrival of the feed truck. Hicks *et al.* (1989) states that generally the highest percentage of animals observed eating in a pen coincides with the time of feed delivery, thus pen checkers often suspect morbidity based on identification of animals that do not feed at this point in time. Because behaviour is such a difficult variable to measure (Parsons *et al.*, 2004), subtle changes in feeding behaviour may go unnoticed until they become more severe. Frequently, the animal only receives medical treatment once it exhibits obvious signs of abnormal behaviour and signs of physical deterioration. The likelihood of successful treatment is highly dependent on the administration of therapeutic drugs early in the disease process. In fact, we now know that the time at which a treatment is first administered is a better predictor

of outcome than the type of drug used or any other factors examined (Blezinger, 2005). Paradoxically however, studies have shown that the currently used methods of treating cattle for BRD are not adequate to prevent production losses, and that improved methods of diagnosis for BRD are needed. It is speculated that with the introduction and hybridization of computer science and artificial intelligence with animal science, subtle differences in feeding behaviour could be detected using automated computer models.

2.1.4. Feeding Behaviour

Of the many individual animal characteristics, and environmental and management factors associated with altered feeding behaviour, health status is recognized as an important but ill-defined contributor (Broom, 2006). In the past, methods used to help researchers understand effects of feeding management on individual animal feeding behaviour have relied on the feeding and monitoring of individually housed animals. Unfortunately, the setup of these former methods influenced and modified animal feeding behaviour when compared to cattle housed under typical commercial conditions (Schwartzkopf-Genswein et al., 2000). As these individually housed animals lack social interactions, it is clear that the information gained under these conditions is unlikely to be relevant to a commercial feedlot. The ability of researchers to observe feeding patterns and their correlation with animal performance has been recently improved with the availability of a newly developed technology, an automatic feed bunk monitoring system (GrowSafe™ Systems Ltd., Aridrie, AB) (McAllister *et al.*, 2000). Through the use of this equipment it is now

possible to accurately monitor feeding behaviour of individual animals within a group or pen without altering their feeding behaviour (DeVries *et al.*, 2003; Gibb *et al.*, 1998; Parsons *et al.*, 2004; Sowell, 1998). This technology has the capacity to monitor feeding behaviour with a degree of sensitivity which allows the detection of feeding behaviour of individuals to be defined within groups of cattle. Using simple measurement techniques, such as feeding behaviour collected with the GrowSafe™ system and presence or absence of metritis post calving, Urton *et al.* (2005) showed that reduced time at the feeder can be used to identify dairy cows at risk of metritis (inflammation of the uterus, a disease common to cows following calving). Although a relationship between feedlot cattle health status and animal feeding behaviour exists (Daniels *et al.*, 2000; Loforgreen, 1983; Parsons *et al.*, 2004; Sowell, 1998), the intricacy of these connections remains unknown and therefore will be further studied in this project. For example, previous research showed that morbid and healthy cattle have different feeding behaviours (Blezinger, 2005; Galyean and Hubbert, 1995). Experts suggest that differences in feeding duration and the number of daily feeding bouts may be the key signs of cattle morbidity. As proof, Sowell *et al.* (1999) recorded severe neophobia (fear of new things or experiences) experienced by presumably healthy cattle during the first four days of the receiving period. They also found that light-weight calves that became sick during the first 32 days after arrival to the feedlot spent 52 % less time at the bunk than presumably healthy calves during the first four days after arrival. These same calves spent an average of 23 % less time at the feed bunk over the initial 32 days following arrival compared to the presumably healthy calves. These findings are consistent with the findings of other studies (Daniels *et al.*, 2000; Parsons *et al.*, 2004; Schwartkopf-Genswein *et al.*, 2005). It has been concluded that cattle feeding behaviour tends to

follow a diurnal pattern (Hicks *et al.*, 1989; Stricklin and Kautz-Scanavy, 1984). This discovery was one of the motivating factors behind Quimby's (2001) work, which led him to suggest that with the use of the GrowSafe™ system, potentially morbid animals may be identified 3-4 days earlier than calves identified via conventional observation via pen checkers.

2.1.4.1.GrowSafe™ System

The need for individual monitoring of feedlot cattle from a physiological perspective arises from the nature of the difficulties involved with monitoring cattle feeding behaviour and animal sickness within a pen. Former methods of animal feeding behaviour observations included tedious, labour intensive, manual methods of monitoring (Streeter *et al.*, 1999). With the introduction of the GrowSafe™ System, detailed feeding behaviour data could now be collected automatically, 24 h a day.

The GrowSafe™ System is modular, and consists of several components. There are two most common variations of the system installed in research institutions, which include a behaviour monitoring system and the feed intake system. Although the underlying concept is similar, the two systems do differ in hardware design and implementation, as well as data collection and processing procedures. The behaviour monitoring system continuously monitors individual feeding behaviour of animals feeding in a commercial environment. It consists of radio frequency identification (RFID) ear tags containing a passive transponder (Figure 2.1), a capacitor, an antenna, a reader panel and a personal computer for data collection. The antenna is incorporated into a rubber mat, which lines the interior surface of the feed bunk. When the

transponder (attached to an animal) comes within 50 cm of the antenna, the reader panel reads the unique transponder number, and sends the data to the computer where it is stored (McAllister *et al.*, 2000). Scanning time is system dependent and varies from 1 to 6.3 seconds.

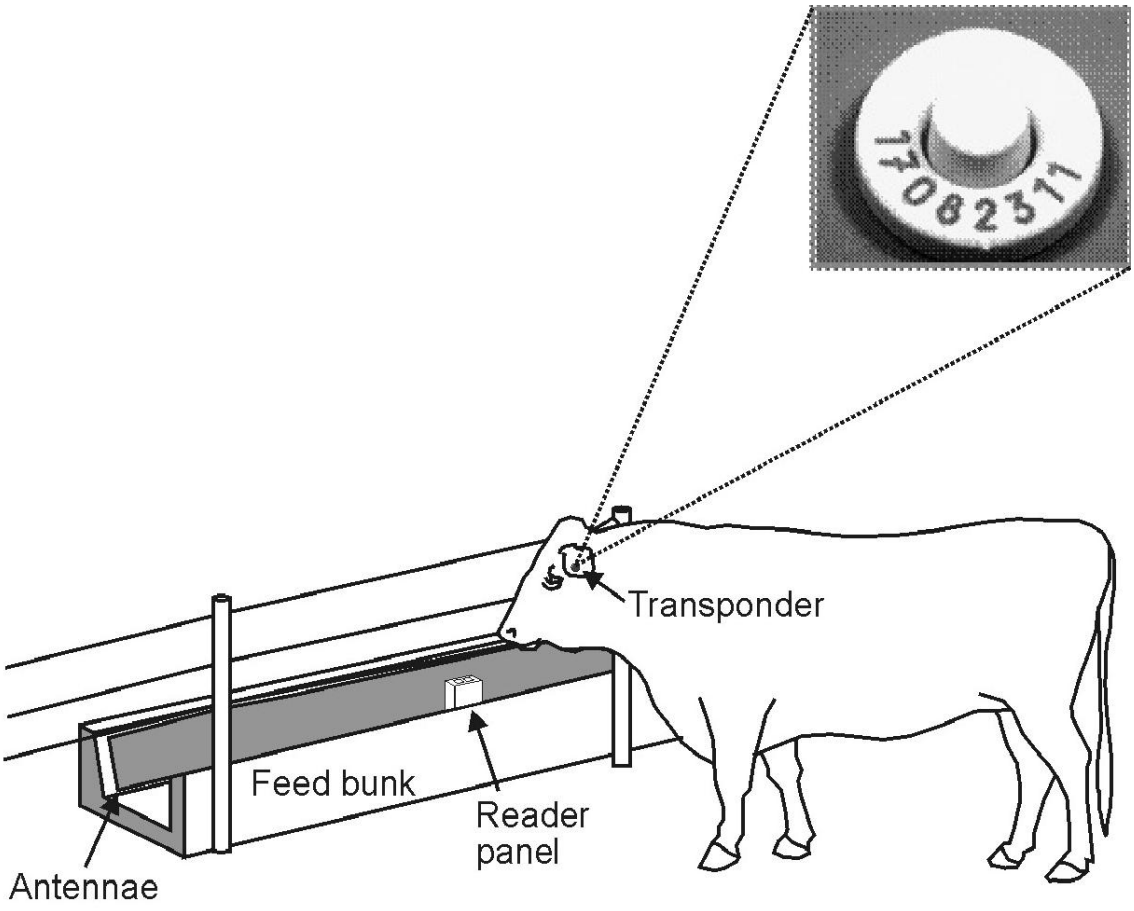

Figure 2.1. GrowSafe™ System

The system is capable of recording an animal's RFID number along with the time that the animal was present at the feed bunk (McAllister *et al.*, 2000). This information is then compiled to determine the duration each animal spends at the feed bunk, and the number of visits made to the feed bunk.

2.1.4.2. Measures of Feeding Behaviour

The introduction of such automated feeding behaviour collection systems allows for the direct measurement and observation of cattle feeding patterns. Detailed measurements taken by these systems are in turn challenging our understanding of the relationship between health status and feeding behaviour, leading to new theoretical constructs and calling old ones into question. To date, there has been little agreement as to which measures of feeding behaviour are most repeatable and valuable when defining feeding behaviour differences between healthy and morbid cattle. Tolkamp *et al.* (1998) and Keyserlingk *et al.* (2002) suggest that animals typically eat in a series of bouts, and this information is often useful to separate the times between events (transponder readings) into within vs. between bouts. Previous research involving feeding behaviour has been based on defining meals in terms of such feeding bouts (Basarab *et al.*, 1997b; Schwartzkopf-Genswein *et al.*, 1999; Sowell, 1998). Results from various bout analysis techniques were incorporated to specify a meal criterion for feedlot cattle as a 300s interval between events that separate within- and between-meal intervals (Schwartzkopf-Genswein and McAllister, unpublished data). This same meal criterion of 300 seconds (where inter-meal intervals must exceed 300S for eating events to be set as 2 different meals) was confirmed in an additional experiment by Gibb and McAllister (1999), where meal length was determined by visual observation of the cattle. Sowell *et al.* (1999) concluded similar results. Measures of feeding behaviour which have been recorded and/or calculated in the past with the GrowSafe™ behaviour monitoring system for research purposes include daily feeding duration, number of meals consumed per day (visits) (McAllister *et al.*, 2000), and inter-meal intervals. Further research involving

feeding behaviour data must be based on the clear understanding of how these measures are currently defined and are summarized in Table 2.1.

Table 2. 1. Definition of feeding behaviour variables. These variables are derived by applying calculations to the raw data obtained from the GrowSafe™ system.

Term	Definition
Feeding event - measured in seconds or minutes	The time interval between the initial detection of the animal's transponder at the feed bunk and the last consecutive reading.
Number of meals or visits – calculated over a specified length of time, such as hour or day.	The number of time intervals between the initial detection of the animal's transponder at the feed bunk and the last time the transponder was detected by the antennae, such that the time between the last two recorded readings was greater than 300 seconds (Basarab *et al.*, 1997b; Schwartzkopf-Genswein *et al.*, 1999; Sowell, 1998).
Daily feeding duration (min d^{-1})	The sum of meal durations during a day. A meal spanning midnight was partitioned based on time in each day.
Inter-meal interval – measured in seconds or minutes	Duration between meals.

Previous research mostly resulted in feeding behaviour observations based on groups of animals. For example, it was observed that animals exhibit a diurnal feeding pattern (Streeter *et al.*, 1999; Stricklin, 1986), as this pattern exists independent of feed delivery times (Schwartzkopf-Genswein, 2003). Hahn (1995) also indicated that cattle

feeding behaviour is influenced by weather and environmental conditions (Johnson, 1985) such as ambient temperature, relative humidity, barometric pressure and wind speed. In a 32-d trial, Sowell *et al.* (1999) concluded that the total time spent at the feed bunk over a 32 d feeding period, was greater ($P< 0.0001$) for healthy than for morbid calves. Forbes (2003) noted, however, that although these patterns of group feeding behaviours emerge, they are a result of combined, distinct and individualistic behaviours. This hypothesis has not been subject to rigorous scrutiny in terms of experimentation, testing and peer review.

2.1.4.3. Calculated Variables and Dataset Setup

Behavioural data can be summarized in various ways, using various techniques. The data summarization process bundles the collected raw data into pre-defined, time-interval data points. Summarizing large datasets can be a challenging task, often requiring expert advice and extensive investigation to identify procedures that are most appropriate for the dataset. The processing is usually assumed to be automated, and typically is unique to the problem. For example, in the case of cattle feeding behaviour, specific summaries are needed when investigating feeding patterns in order to maximize our understanding of differences in feeding behaviour between healthy and sick cattle. A compact summary of the data can be obtained by processing the data by day, where unique values of feeding behaviour measurements would be assessed over the duration of a day. As an example of the most extreme capabilities of this system, it is possible to summarize the data by the minute. Although this approach is very precise and would highlight even the smallest inter- and intraday differences in animal feeding behaviour, it

would require very large storage space and would be unfeasible because of data handling and time constraints.

2.2. The Conjunction of Animal Science and Computer Science

In the field of pattern recognition, the prediction of behavioural patterns over time is usually based on some historical knowledge of "normal" behaviour that is used as a standard of comparison for changes in behaviour. For example, in the case of cattle feeding, pen checkers often associate repeated absence of the animal from the feed bunk during feeding time, with poor health (Edwards, 1980). Thus, when considering a proposed computer model for identifying feeding behavioural anomalies within a feedlot pen, the challenge is to build a system that is able to consider the normal diurnal fluctuations in behaviour of all animals as being distinct from those behaviours that are indicative of morbidity.

2.3. Artificial Intelligence (AI)

With the introduction of the digital computer in the twentieth century, AI became a viable discipline. It is a field of computer science, concerned with the automation of intelligent behaviour (Luger, 2002). One of the pioneers of AI was the British mathematician Alan Turing, who gave the first scientific discussion of human-level machine intelligence. He is well known for his contributions to the theory of computability and several inventions. These include the Turing Machine, a simple

abstract computational device intended to help investigate the extent and limitations of what can be computed as well as the Turing Test, in which the performance of a presumably intelligent machine (Turing Machine) is measured and compared against human intelligence (Kak, 1996). Currently, there are many tasks that humans can perform that can not yet be performed by a computer. In contrast, some complex mathematical calculations and formulas that can easily be solved via a computer are too complicated for humans to process in a timely fashion. McCarthy (1996) argues that reaching human-level AI requires programs that deal with common sense informative circumstances, in which the phenomena to be considered in achieving a goal are not preset. For example, the concept of "recognition" seems simple and familiar to most people. Recognizing a specific object or well-defined behaviour is a task humans frequently and commonly perform. However, recognizing behaviour in terms of datasets and numerical values presents a far greater challenge for humans and can be far more easily accomplished through computational theory. One of the branches of AI that is concerned with the identification of behaviour and studies the operation and design of automated decision systems is pattern recognition.

2.3.1. Automated decision systems

The most salient characteristic of automated decision systems is that they actually make a decision. In many cases their decisions are made without any human intervention at all, in others – sometimes for legal or ethical reasons – they work alongside a human expert such as a doctor. The intention of the following sections is to

give an overview of how automated decision system models are created, from both a biological and a mathematical perspective.

2.3.1.1. Pattern Recognition

Pattern recognition is defined as the process of identifying structure in data by comparison to known structures (Dutta and Dutta, 2006). Today, as data are being collected and accumulated at a dramatic rate and the availability of large databases is intensifying, demands on automatic or semi-automatic pattern recognition systems are on the rise. Watanabe (1985) defines a pattern "as opposite of chaos; it is an entity, vaguely defined, that could be given a name." For example, a pattern could be a face, sound signal, a fingerprint image, or feeding behaviour. The aim of pattern recognition systems is to associate each pattern with existing pattern classes (Dutta and Dutta, 2006). The key to most pattern recognition systems however, is abundant good-quality data.

2.3.1.2. Data Acquisition and Quality

Data can be collected by various means, for example through experiments, observations, theory, models and simulations. In the past, data usually were presented as tables of numbers but presently scientific data are most often stored in databases and can involve numbers, text, images, diagrams, pictures, and equations. Efficient methods of data acquisition are fundamental to the generation of the extensive datasets that are required to define complex behaviours. In many cases, sensors transduce physical conditions into electrical signals that can be digitized and stored for subsequent computer analysis. Dedicated instrumentation makes it possible to collect detailed

observations on an immense scale, and advanced electronics and computers have simplified some experimental operations and made processes such as repeat measurements less labour intensive. The cattle-feeding behaviour information collected via the GrowSafe™ system is an excellent example of such an application. Although collected data are more precise and detailed as compared to data collected using former methods, data quality control must still be implemented in order for the system to be viable and pragmatic. Problems with data quality may stem from various sources, including system deficiencies, loss of signal, and malfunction of the system. In other words, the quality of the data often depends on the design and production process involved in generating the data. While most errors in data within these systems are often barely observable, the cumulative impact of poor data quality on final interpretation of the dataset can be enormous.

2.3.1.3. Data Quality

The subject of data quality has been addressed in several research areas, including statistics, accounting, management, and computer science. It has been defined in several ways in the literature. For instance, Orr (1998) describes it as "the measure of the agreement between the data views presented by an information system and the same data in the real world", whereas other definitions refer to a set of dimensions such as accuracy, completeness, consistency and timeliness (Ballou and Pazer, 1985).Wand and Wang (1996) explicitly give 5 dimensions for defining data quality: accuracy, completeness, consistency, timelessness, and reliability. Wang and Strong (1996), elected to select 15 different dimensions to be the most important out of an initial 179.

Thus, even though some dimensions are considered to be universally important, scientists do not agree on a single set of dimensions as being unanimously important in assuring data quality. Wand and Wang (1996) also suggest that the notion of data quality depends on the actual use of the data. In particular, what may be considered good quality data for a specific application may not be of adequate quality for other purposes (Ballou and Tayi, 1999). For instance, at the feedlot feed intake data collected on a daily basis is sufficient when calculating the amount of feed to be delivered by the feedtruck, whereas the quality of such data would prove to be poor when attempting to define the feeding behaviour of individual cattle within the pen throughout the day. Different users have different data quality requirements. Consequently, it is important to provide a design-oriented definition of data quality that will reflect the intended use of the information and will lead to input datasets that are of satisfactory quality when employed in a pattern recognition system (Wand and Wang, 1996).

2.3.1.4. Knowledge Discovery in Databases (KDD)

Fayyad *et al.* (1996) describes the flourishing field of knowledge discovery in databases, also referred to as data mining, as a powerful method and technique for interpreting data. This process has been applied to many domains including astronomy, marketing, investment, manufacturing, fraud detection and scientific research (Fayyad *et al.*, 1996). As described by Fayyad *et al.*, (1996), KDD can be defined as a structured, interactive and iterative process, involving several steps with many decisions made by the user, namely:

1. developing an understanding of the application domain,

2. creating a target dataset,
3. data cleaning and preprocessing,
4. data reduction and projection,
5. matching the goals of the KDD process to a particular data-mining method,
6. exploratory analysis and model and hypothesis selection,
7. data mining,
8. interpreting mined patterns and
9. acting on the discovered knowledge.

Items 1 – 4 and 7 will be discussed in detail, as they are most relevant to this thesis.

Developing an understanding of the application domain

It is crucial to understand the input to any pattern recognition system and to know the strengths and weaknesses of the input prior to the knowledge-discovery process. This knowledge can be obtained from manuals, domain experts, and literature. In the case of cattle feeding, it is imperative for the dataset to represent a true reflection of cattle feeding behaviour in a typical feedlot environment as described in the Animal Science section of this literature review. It is also important to recognize and take note of errors and problems such as system malfunction during data acquisition, as some of the difficulties that arise in the pattern recognition process often depend on the quality and limitations of the input data. Understanding the sources of error and limitations and why they are important, is key to the development of a robust pattern recognition algorithm.

Creating a target dataset

As some data collection systems result in abundant data, data mining experts suggest reducing the dataset in size to effectively meet the needs of the analysis. This is

achieved by eliminating redundant or irrelevant data and creating a sub-dataset that consists of information that is most intrinsically interesting and relevant to the test hypothesis. Just as insufficient data to a system would yield poor results, too much or excess information would also clog the system, and may result in poor output. Domain knowledge is beneficial for intelligent reduction of the dataset, as it requires the user to make knowledgeable decisions.

Data preprocessing and data cleaning

When given a poor description of an object, humans often will incorrectly identify it. Similarly, poor data quality can lead to incorrect interpretations no matter how robust the pattern recognition algorithm (Redman, 2004). As is the case with humans, in an automated recognition system, the process depends greatly on the quality of the information provided. Gaining new information and knowledge of a specific domain depends largely on data analysis. However, data analysis is only efficient if the datasets provided for analysis are error free. Often the efficiency and effectiveness of data analysis is hampered by data anomalies (errors), making the identification of existing or potential problems in poor quality datasets important in terms of data processing, which usually involves cleaning the data before data mining tools are applied. Thus, preprocessing of the data, also referred to as filtering is key to a solid and robust pattern recognition system (Fayyad *et al.*, 1996).

During the data preprocessing phase the data are transformed into a format that is usually more easily and effectively processed via an analysis process such as pattern recognition. It seems however that there are no general guidelines as to how to determine

the appropriate data pre-processing techniques. Famili *et al.* (1997) describes a specific transformation (T) in terms of the raw real-world data vectors X_{ik} and Y_{ij}:

$$Y_{ij} = T(X_{ik}) \quad (2.1)$$

where Y_{ij} is the newly created dataset that preserves the 'valuable information' in X_{ik} but eliminates at least one of the problems in X_{ik} ,

i=1, . . . n where n = number of objects,

j=1, . . . m where m = number of features after preprocessing (generally $m \neq l$.) and

k=1, . . . p where p=number of attributes/features before preprocessing.

Famili *et al.* (1997) also discusses two main reasons for performing data preprocessing:

1. to fix problems that may arise with the data and
2. to prepare the data for analysis.

There are several unique preprocessing techniques described by Famili *et al.* (1997), among which data cleaning/filtering is described under the data transformation section.

There are often many problems with real-world data. Cleaning these data is a time consuming task, as any errors and inconsistencies in the dataset must be identified and then addressed. Data cleaning is a term without a precise or fixed definition, perhaps due to the fact that it is domain dependent and application specific (Maletic and Marcus, 2000; Mathieu and Khalil, 1998). Current data cleaning methods do exist, and focus mainly on the transformation of the data and the elimination of duplicates in a dataset (Famili *et al.*, 1997; Maletic and Marcus, 2000). Missing values may often impose great concern, as missing data resolution can be a challenge and may present compelling research problems such as predicting preterm birth risk patterns as described by Grzymala-Busse *et al.* (2005). The removal of unwanted information or data from the input is application dependent, thus the filter algorithm or method to be implemented is

usually unique to the project and demands extensive domain knowledge so that useful information is not lost (Maletic and Marcus, 2000).

Data reduction and projection

As machine learning aims to tackle larger, more intricate tasks, data reduction becomes an imperative step toward understanding and discerning distinct patterns from large and complex datasets. Patterns are typically described in terms of multidimensional data vectors, where each component is called a feature (Duda *et al.*, 2001) . The process where the dimensionality of the dataset is reduced to a set of more vital features is called feature extraction. The objective of feature extraction is to characterize an object to be recognized by measurements whose values are very similar for objects in the same category, and very different for objects in different categories. It is a process of studying and deriving useful information from filtered input patterns and identifying the most effective subset of the original features to later use in the classification process. This approach leads to the smallest classification error. The methods of feature extraction and the extracted features are application dependent; however Blum and Langley (1997) classified feature-extraction techniques into three basic approaches:

1. embedded approach: features are added or removed in response to prediction errors of a simple embedded classifier,
2. filter methods: methods work independently to remove features without knowing the effect on the classification algorithm (Principal Component Analysis (PCA) is an example) and
3. wrapper methods: evaluate candidate feature sets using a classification algorithm on the training data.

Feature extraction is often regarded as dimensionality reduction. One way to reduce the dimensionality of the dataset is by identifying major factors behind the variability of all variables, through the means of PCA (Section 2.4.1.6).

Data mining

Many theories and algorithms have been proposed and studied extensively for understanding and summarizing data, and deriving knowledge from data. The spectrum ranges from classical analysis, cluster analysis, and data analysis to recent machine learning, data mining, and knowledge discovery. One of the main goals of data mining is to provide a comprehensible description of information extracted from databases. Given a pattern, the act of recognition and/or classification can be divided into two broad categories (Scott, 2006):

a. supervised classification – where the input pattern is recognized as a member of a predefined class and
b. unsupervised classification – where the pattern is assigned to a previously unknown class.

2.3.1.5. Classification

Data classification is the final stage of pattern recognition. This is the stage where an automated system declares that the presented object belongs to a particular category. There are many classification methods in the field, including:

1. member-roster concept – an input pattern is compared with sets of patterns stored in a classification system and placed under the matching pattern class,

2. common property concept – the properties of an input pattern are compared with properties of patterns stored in a classification system, and the pattern/object is placed within a class which has similar common properties and
3. clustering concept – input patterns are presented as vectors and the relative proximity to representative cluster vectors is used to classify patterns within the target classes. If the target vectors are distinct, i.e. far apart in a geometrical arrangement, it is easier to classify the unknown patterns. Subtle differences in the classes are characterized by vectors that are nearby and more complex algorithms are required to classify the unknown patterns. Minimum-distance classification is one simple algorithm, which computes the sum of squared differences between the unknown pattern and the representative patterns for the clusters. The unknown pattern is assigned to the class that results in the least sum. This algorithm works best when the target patterns are easily differentiable.

The conceptual boundary between feature extraction and classification is somewhat arbitrary; an ideal feature extractor would yield a representation that makes the job of the classifier trivial; conversely, an omnipotent classifier would not need the help of a sophisticated feature extractor.

2.3.1.6. Principal Component Analysis (PCA)

PCA was originally introduced in 1901 by Karl Pearson – who defined it as a mathematical method to achieve dimensionality reduction, as it consolidates redundant data and condenses essential information into fewer variables (Lavine, 2005). The

underlying goal of PCA-based dimension reduction is described in terms of dimensionality reduction of the dataset as a liner transformation. This technique provides an optimal way of reducing dimensionality by projecting the data onto a lower dimensional orthogonal subspace that captures as much of the variation of the data as possible. PCA transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called principal component (PC)), the second greatest variance on the second coordinate, and so on. It is well documented, that lower order PCs often contain the "most important" aspects of the data (Lavine, 2005). PCs are distinct, and comprise the variability of the dataset; and are sorted in order of significance of variance among all PCs (Lavine, 2005). By selecting the minimum number of PCs that capture most of the variation in the dataset, a 'subspace' (defined as more suitable for data visualization and analysis than the original dataset) can be identified. It is common practice to apply a K-means clustering technique to the chosen PCs (Ding and He, 2004).The field of pattern recognition and classification outlines numerous clustering algorithms such as K-means clustering (Duda *et al.*, 2001). The choice of the most appropriate method depends on the specific nature of the problem. Yeung and Ruzzo (2001) emphasize however, that clustering with the PCs rather than with the original dataset enhances cluster quality only when the right number of components or when the right set of PCs are chosen.

2.3.1.7. K-means clustering

A cluster of objects is most commonly defined in terms of their similarity to one another. Similarity is usually measured by a distance function defined on pairs of data

points. There is a variety of ways to calculate distance, with the Euclidean distance calculation being the most common method (Chang, 2007; Wang, 2006). The data must be normalized before K-means clustering is applied, as larger scaled variables can dominate others, resulting in skewed results. In pattern recognition, it is common practice to apply K-means clustering techniques to data that have been reduced in dimensionality via PCA. In fact, Ding and He (2004) demonstrated that PCs are a continuous solution to the discrete cluster membership indicators for K-means clustering. The K-means algorithm is a process used to cluster objects based on given attributes into K partitions or clusters, such that intra-cluster variance is minimized, whereas inter-cluster variance is maximized. A cluster is defined as a group of objects with similar features (Duda *et al.*, 2001).The goal of K-means clustering is to divide the data points into K clusters such that some metric relative to the centroids of the clusters is minimized (Chang, 2007). A centroid is defined as the mean of all data points already assigned to a cluster – thus each cluster has a centroid. Initially, K random points from the dataset are selected. These points represent initial cluster centroids. New data points are assigned to a cluster based on the estimation of Euclidean distances between it and each centroid. The new data point is assigned to the closest cluster and the new centroid, once the newly assigned member is taken into account, is defined as the updated mean of the new cluster. This procedure is iterated until the centroids no longer change, resulting in the separation of the original dataset into K distinct clusters (Chang, 2007).

The analysis of feeding behaviour patterns from the viewpoint of determining abnormalities (sickness) has great bearing on the feedlot industry (Hickman *et al.*, 2002). The data collected with the GrowSafe™ system is comparable to that of other data

received in signal processing experiments. Although the concept of signal processing is not novel, and has been applied in various fields such as sound, and image and character recognition, tailoring some of the ideas to fit cattle feeding behaviour data is original.

2.4. Summary

The objective of this work was to develop and test a classification process using pattern recognition techniques that would identify morbid feeding behaviour prior to the animal exhibiting physical signs of sickness.

3. IDENTIFYING CATTLE SICKNESS EARLIER THAN TRADITIONAL METHODS USING PATTERN RECOGNITION TECHNIQUES

In Chapter 3 the development of a pattern recognition process is introduced in the form as presented in scientific journals. First, emphasis is given to animal health status definitions, followed by a precise data cleaning process. Feeding behaviour is summarized by processing the raw data by 4-h time intervals. Data mining and pattern recognition techniques were applied to the variables to conclude the health status of individual animals. The performance of the developed process is presented in the results section of this chapter. The discussion compares the work presented in this chapter to previously reported research.

3.1. Introduction

Bovine Respiratory Disease (BRD) is one of the most prominent and economically important diseases of feedlot cattle (Duff and Galyean, 2007; Smith, 1998). It is a disease of the respiratory tract, caused by stress, viral and bacterial infections, and numerous other stressors and agents such as dust, cold and fatigue (Bagley, 1997; Duff and Galyean, 2007; Griffin, 1998; Loerch and Fluharty, 1999). BRD is of significant concern to feedlot operators in terms of animal welfare and economic loss (Duff and Galyean, 2007; Loneragan, 2001), accounting for 65-77 % of feedlot cattle morbidities and 44-72 % of mortalities in the United States (Edwards,

1996; Galyean *et al.*, 1999; Quimby, 2001). Approximately 65 to 80 % of cases of BRD occur in cattle during their first 45 days at the feedlot (Griffin, 1998; Mathison, 1993; Smith, 1998). Physical signs of BRD include thick nasal discharge, elevated temperature, difficulty breathing, discharge from eyes, red peeling muzzle and listless behaviour (Galyean *et al.*, 1999; Griffin, 1998).

One of the key indicators of morbidity feedlot personnel use to identify potentially sick animals is animal behaviour with particular emphasis on feeding behaviour (Broom, 2006). Visual observation is still one of the most reliable methods of identifying morbidity in feedlot cattle (Duff and Galyean, 2007). However, subtle changes in behaviour may go unnoticed until the animal shows obvious clinical symptoms at an advanced stage of the disease.

In the past, analysis of animal behaviour has been an arduous task, requiring a human observer to record and classify individual actions. The development of an automated bunk monitoring systems allows for the collection of detailed cattle feeding behaviour data 24 h a day on all cattle within a pen. Several studies using similar automated systems have reported significant differences in the feeding behaviour of healthy and morbid cattle. Quimby et al. (2001) found that using cumulative sums analysis (CUSUM; SAS institute, Inc. 1995), morbid animals could be identified up to 4.1 days earlier than by a pen rider using visual observation as a determinate of health status. Daniels *et al.* (2000) reported that morbid calves spent 40 to 41 % fewer minutes per day at the feed bunk than untreated and presumably healthy calves over two 21-d receiving trials. All of these studies used simple linear statistics to compare feeding behaviour parameters such as bunk attendance duration and frequency.

Non-linear data mining analysis techniques such as clustering, machine learning procedures and algorithms have been previously used to identify patterns in biological data. For example, these techniques have been employed to assist radiologists to identify and classify types of mammary tumour lesions (Masala, 2006). Classification algorithms and methods such as neural networks, Bayesian networks and genetic algorithms have also been employed in the detection of patterns in biological data. Application of non-linear methods on detailed feeding behaviour data may be useful in identifying different patterns of behaviour between healthy and morbid cattle. To date, no studies have used non-linear methods such as pattern recognition to analyze feeding behaviour in an attempt to identify morbid animals. The objective of this study was to develop an algorithm applying pattern recognition techniques to data on individual feeding behaviour to enable earlier detection of morbidity in feedlot cattle than conventional methods.

3.2. Materials and Methods

Two groups of animals were used in this study. Data from one group was used to form the model dataset, whereas data from the other group formed the naive dataset.

3.2.1 Animals (Model Dataset)

Three hundred and eighty-four (384) non-preconditioned, predominantly British x Continental heifers, averaging 228 ± 22.7 kg (initial BW) were monitored over a 225 d feeding period in four separate feedlot pens at the Cactus Feeders feed yard in Amarillo,TX. The number of steers assigned per pen was adjusted to provide each animal with approximately 24 cm of bunk space and 14 square meters of pen space at the beginning of the tiral. The pens were equipped with the GrowSafe™ (Airdie, AB)

feed bunk monitoring system. Heifers were purchased from one of two auction markets (Wilson, TX and Meridian, MS) and were transported a distance of 580 or 950 km, respectively to the feedlot on January 19, 2002, where they were processed and held in receiving pens for 2 days before they were randomly allotted to their home pens. At processing heifers were administered Micotil ™ (Elenko, Greenfield, IN) and given a Synovex-H ™ (Wyeth Animal Health, Guelph, Ontario) implant and were re-implanted 115 days later using Finiplex-H™ (Intervet Animal Health Inc., Boxmeer, The Netherlands). Cattle were adapted to the finishing diet (Table 3.1) using a two-ration system that incorporated the feeding of the basal starting diet (approximately 36 % roughage on DM basis) and the basal finishing diet (approximately 9 % roughage on a DM basis). During transition to the final finishing diet, all pens were fed three times daily through a series of 10 feeding phases that progressively increased the energy content of the diet. Cattle completed the final feeding phase and were on the finishing diet after approximately 45 days. After the transition period was completed all study pens were fed three times daily at approximately 0600, 0900 and 1300 h. In the third feeding, MGA (Pfizer Animal Health) was fed to provide heifers 0.5 mg/hd/day. Basal diets were prepared in the feed yard mill, which was equipped with a computerized batching system and horizontal paddle mixer. Diets were formulated to meet or exceed National Research Council (1996) requirements for growing - finishing beef cattle. Carcass information as well as incidence of lung lesions and liver abscess were collected on all animals at the time of slaughter.

Table 3. 1. Composition of basal diets, dry matter basis for Model Dataset

Item	Diet	
	Starting	Finish
Ingredient		
Steam-flaked corn	53.1	56.4
High-moisture corn	---	21.1
Alfalfa hay, chopped	33.6	4.2
Corn silage	6.6	6.7
Animal fat	---	3.8
Liquid starter supplement	6.7	---
Finisher supplement	---	7.8
Additives[b]		
Monensin, g/ton	15.3	32.4
Tylosin, g/ton [b]	0.0	9.3
Vitamin A, IU/lb.	3,600	2,258
Vitamin D, IU/lb.	360	226
Vitamin E, IU/lb.	20	5
Calculated Composition		
Dry matter, %	70.90	71.06
NEm, Mcal/100 lb.	82.66	99.66
NEg, Mcal/100 lb.	54.01	69.10
Crude protein, %	14.00	13.50
NPN, %	2.25	3.30
Crude fat, %	3.44	7.46
NDF, %	23.31	12.62
Calcium, %	0.85	0.55
Phosphorus, %	0.44	0.30
Magnesium, %	0.26	0.20
Potassium, %	1.40	0.65
Sulfur,%	0.21	0.19

Melengestrol Acetate (MGA) fed in third feeding of finishing diet to provide 0.5 mg/hd/day

3.2.2. GrowSafe™ System

Individual feeding behaviour was collected with GrowSafe™ 24 h a day over the 225 d experimental period. The GrowSafe™ system has been previously described in detail by (Parsons *et al.*, 2004; Schwartzkopf-Genswein *et al.*, 1999). The system

consisted of five panels in which antennae were embedded in a rubber mat that lined the entire length of the feed bunk of all 4 feedlot pens used in this study. As illustrated in Figure 3.1 each pen was monitored by more than one panel.

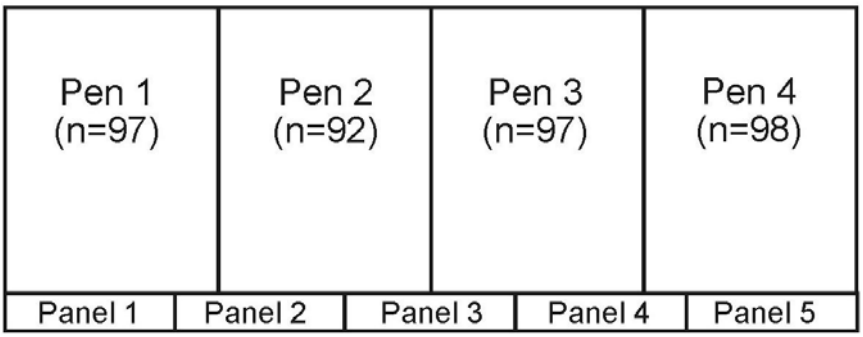

Figure 3.1. Layout of GrowSafe™ system panels and distribution of animals in each pen.

Each panel functioned independently, limiting system failure to faulty panels only. The raw data collected by the system consisted of the unique transponder number assigned to an animal, a Julian date and time stamp, and a location along the feed bunk where the animal was feeding. This information was later processed and summarized to generate new variables as later described in detail in Section 3.2.4. As the system hardware was exposed to harsh physical and environmental conditions, malfunctioning of the system did occur, resulting in some lost data. Radio frequency systems are known to be vulnerable to interference from a multitude of sources, such as equipment or metal surrounding the antennae (Schwartzkopf-Genswein *et al.*, 1999) which may also contribute to error in data acquisition.

3.2.2.1. Sync Chip

In an effort to confirm the validity of the collected information, GrowSafe™ hardware incorporates a sync chip whose purpose is to identify when the system is not

functioning. The sync chip is an RFID transponder that was embedded into the GrowSafe™ panel in close proximity to the antennae and was integral to the data cleaning procedure described later in Section 3.2.6. These data were used to exclude feeding behaviour data collected during periods of time when system functionality was suboptimal.

3.2.3 Health Status Classification

Cattle were defined as morbid (M) if they were removed from their home pens for medical assessment and were treated on one or more occasions at any point over the 225 d trial. Animals were removed for treatment according to the visual observation of experienced feedlot personnel, assessed and diagnosed by staff members and treated accordingly. The type of illness the cattle were being treated for was recorded, and the animal (depending on the severity of sickness) was moved to a hospital pen where it was further monitored and treated or in less severe cases the animal was returned to its home pen post-treatment.

Animals removed from their home pens for sickness were given the M classification; dead and prematurely culled cattle were not included in the M group. Only cattle diagnosed with BRD at the time of treatment were used in the study. Animals that had not been removed for sickness and did not have any lung lesions or liver abscesses at slaughter were subsequently defined as healthy (H). It was assumed that animals having lung lesions suffered from BRD at some point during their lives. Cattle that were never removed for morbidity, but had lung lesions and/or liver abscesses at slaughter (i.e. not healthy) were categorized as having unknown (U) health status.

To reduce the possibility of false positive (categorizing H animals as M) and false negative (categorizing M animals as H) classifications, a severity index of morbidity was created. A unique procedure was introduced where the expert advice provided by animal scientists was combined with each animal's medical record resulting in identifying a measure of confidence in a correct M classification defined as a confidence level of sickness (CLS). The procedure required that all M animals be divided into three subgroups based on the number of time (1, 2 or 3) that they were removed for treatment after observation of morbidity. From an animal's treatment, CLS classifications were assigned that incorporated both the total number of removals and the number of days spent in hospital upon first removal only. Figure 3.2 illustrates the number of days spent in hospital by animals in each subgroup upon their first removal from the pen. The color intensity represents the number of animals falling into the corresponding x-y coordinates (i.e. the darker the point, the more animals). In all three subgroups the median number of days spent in the hospital was 3. This was assumed to be a consequence of the feedlot management practices and protocol required for specific antibiotic treatments.

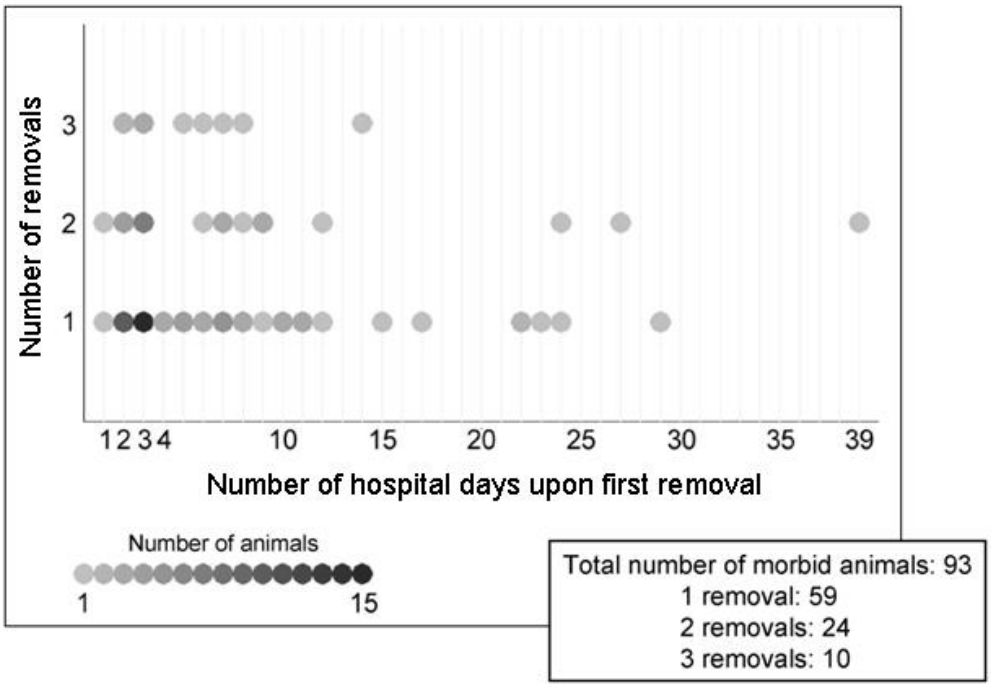

Figure 3.2. Median number of days calves spent in the hospital, upon the first occasion they were removed from their home pen for medical assessment and/or treatment.

From treatment histories illustrated in Figure 3.2, criteria were developed by which to classify the animals according to CLS the following way (Table3.2):

1. Low: identified morbid once and spent up to 3 days in hospital after removal, or identified morbid twice and did not spend time in hospital upon first removal.
2. Moderate: identified morbid once and spent more than 3 days in hospital after removal, or identified morbid twice and spent 1, 2 or 3 days in hospital upon

first removal, or identified morbid 3 or more times, and did not spend time in hospital upon first removal.

3. High: identified morbid twice and spent more than 3 days in hospital upon first removal or identified morbid 3 of more times and spent more than one day in the hospital upon first removal.

Table 3. 2. Strategy to define the level of confidence associated with having been identified as morbid based on number of removals from home pen and days spent in hospital upon first removal.

Days in hospital upon first pull	Number of removals			
	1	2	3	>3
0	Low	Low	Moderate	Moderate
1, 2, or 3	Low	Moderate	High	High
>3	Moderate	High	High	High

3.2.4. Calculating behaviour data variables

The following subsections describe the data processing routine (Figure 3.3) used prior to the application of a pattern recognition algorithm.

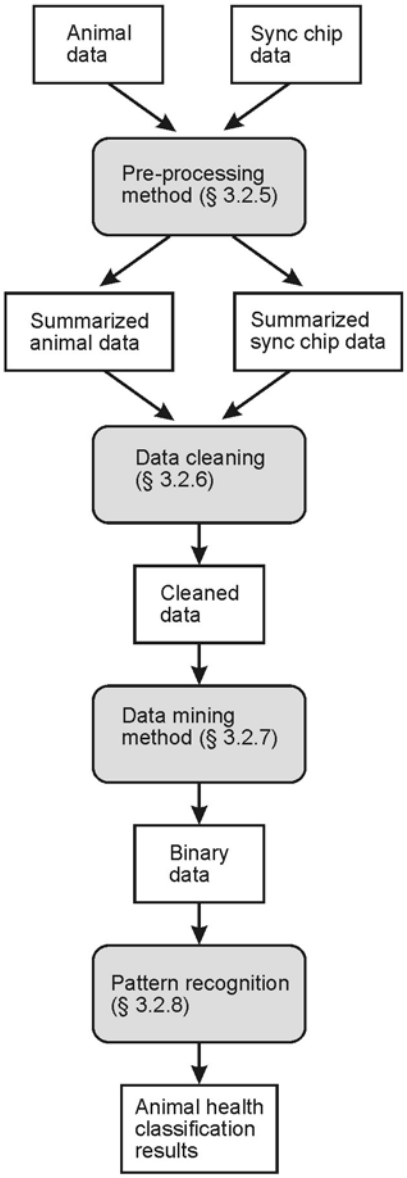

Figure 3.3. Summary of data processing routine.

3.2.4.1. Processing Period

Data collected from the animals in each pen and the five sync chips embedded into each of the five GrowSafe™ panels were stored in raw form in individual binary output files generated by the system onto a personal computer and were processed independently. Data from all animals and each sync chip were summarized into 4 h periods starting at hour 0200 on the first day of the experiment, resulting in six distinct

periods per day as follows: period 1 (0200-0600), period 2 (0600-1000), period 3 (1000-1400), period 4 (1400-1800), period 5 (1800-2200), and period 6 (2200-0200). The 4 - hour processing period was selected based on the differences of feeding and diurnal feeding patterns of M and H cattle, over the 5 d period before M cattle were removed from the pen (Figure 3.4.).

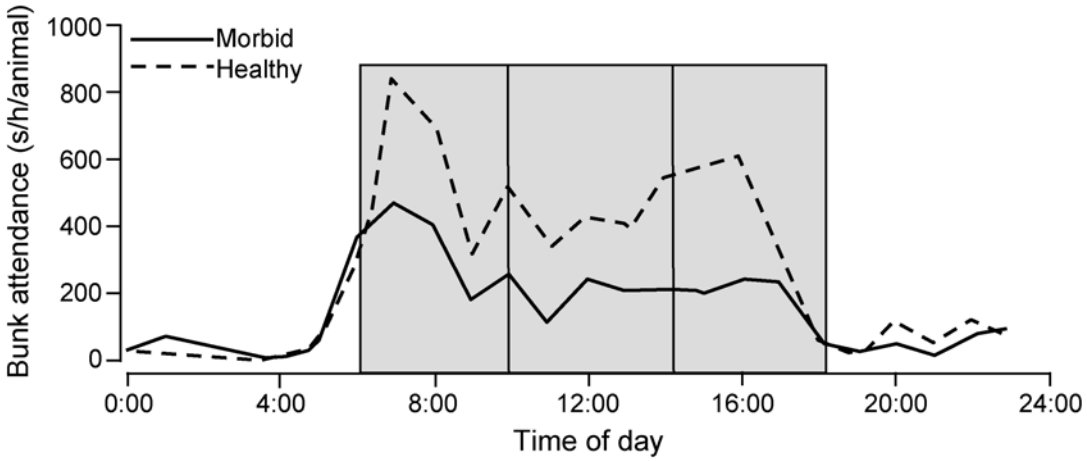

Figure 3.4. Average diurnal feeding pattern of Morbid (M) (n=10) and Healthy (H) (n=10) animals over a 5 d period prior to M cattle being removed from their pen.

3.2.5. Pre-processing Method

One of the initial tasks of the processing routine was to transform the raw information obtained from the GrowSafe™ system into a form that could be more easily interpreted. This was achieved by converting the raw data into text files with the GrowSafe™ software (Version 5.0). The resulting text files consisted of date-time stamps and animal identification number. The text files were then imported into custom software previously developed using Visual Basic 6.0 (Microsoft Corporation, Redmond, WA) combined with an Oracle® based database (IRAD) (Oracle Corporation, Redwood Shores, CA). Data were compiled into a format where the start and end of each

feeding observation were used to define feeding events and their duration in seconds. A feeding event was defined as the length of time an animal spent at the feed bunk without interruption. An interruption was considered the absence of an animal from the feed bunk for a period longer than 5.25 seconds. Several factors may have caused an interruption, including displacement by another pen mate, human interference or the animal simply taking a break from feeding. Feeding events that were separated by an interruption ≤300 s in length were grouped into meals. Meals were separated from each other by interruptions > 300 s in length as previously described by Schwartzkopf-Genswein *et al.* (2002). Interruptions separating each meal were defined as inter-meal intervals. IRAD software was used to summarize the feeding events based on the previous definitions, resulting in a new dataset that contained animal transponder numbers, date-time stamps indicating the start of a meal, and duration of the meal in seconds. This information was further processed and summarized by the 6 time periods previously described in Section 3.2.4.1 from which an additional 11 feeding behaviour variables were derived. The new feeding behaviour variables were calculated from the two core variables, which included feeding duration (dur) and the inter-meal interval (int), where $\sum$ dur + $\sum$ int = length of the processing period. Variables derived included minimum, maximum, average, total and standard deviation of feeding durations and inter-meal intervals as well as the number of meals or visits made to the feed bunk and the number of inter-meal intervals over a 4-hour period, resulting in a total of 12 variables.

An algorithm used for summarizing data by time intervals was also developed to better understand the manner in which individual animals use the feed bunk throughout the day. This algorithm – developed using PL/SQL (Oracle Corporation, Redwood

Shores, CA) - considered all possible combinations of meal lengths and processing period span times. Figure 3.5 and Table 3.3 illustrate and highlight the implemented rules used to derive values for the previously listed feeding variables.

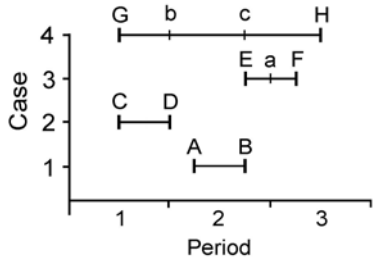

Figure 3.5. Four distinct ways a feeding event may span across successive 4-h periods. Horizontal line segments represent animal feeding behaviour occurrence.

Table 3. 3. Applied calculations of rules implemented for specific examples demonstrated in Figure 3.5.

	Period 1		Period 2		Period 3	
Case	Duration	Visits	Duration	Visits	Duration	Visits
1	0	0	B-A	1	0	0
2	D-C	1	0	1	0	0
3	0	0	a-E	1	F-a	1
4	b-G	1	c-b	1	H-c	1

Four distinct cases are demonstrated: case 1 represents a meal occurring within a given time period (2). In this case, point A represents the beginning of the meal, and point B represents the end, thus the length of the meal is equal to the length of the line segment AB. No meals were recorded for Periods 1 and 3, and one meal was recorded for Period 2 (Table 3.3.). The rules for partitioning feeding bouts and visits were implemented in the following way: Let P1 and P2 indicate the beginning and end times of processing period P respectively, and let P be the set of all p_is such that $P1 \leq p_i < P2$. A data point (time point) t belongs to P if and only if t Є P. Therefore point D (from Case 2), in fact belongs to Period 2. Although point D has no length and therefore no duration (in Period2), the fact that point D exists resulted in a recording of a meal event in

Period 2 as well as in Period 1. Mathematically, the calculations and recordings are correct, but biologically a meal of 0 length has little relevance. This problem was addressed and corrected by introducing a technique that identified such discrepancies. This method involved the removal of the 1-s visit that was assigned to the subsequent period. Case 3 demonstrates when a meal extended over the boundaries of two periods. Here, the meal was divided into two segments, each of which fell into Periods 2 and 3. The length of meal recorded for Period 2 was calculated as the length of line segment Ea, and the length of meal for Period 3 was aF. Case 4 demonstrates a scenario where a meal extends over the entire data collection period, thus a meal was recorded for each processing period. Period 1, 2 and 3 had meal lengths represented by line segments Gb, bc, and cH, respectively. It is important to note that the sum of all meals processed throughout any x-hour period using y-hour processing periods does not necessarily equal the number of meals if that same time frame was processed using a z (where $x \neq z$) hour processing period. In Figure 3.5 for instance, Case 4 identifies three meals in total, over the sum of all three processing periods when the periods were processed individually. However, if the data were processed as one segment, the number of meals calculated would be just one.

3.2.5.1. Inter-meal Interval

Although mathematically the importance of inter-meal intervals appears redundant, biologically it proved to be an important variable, providing information about how the animals fed. To demonstrate this, consider the example illustrated in Figure 3.6, of three distinct feeding behaviour patterns having the same feeding duration. The number of visits and the number of inter-meal intervals separated examples 1 and 2

from example 3. However, other differences in feeding behaviour are highlighted in the minimum and maximum values of the inter-meal intervals, which are not the same in any of the three feeding scenarios presented even though duration in all three examples were identical (Table 3.4).

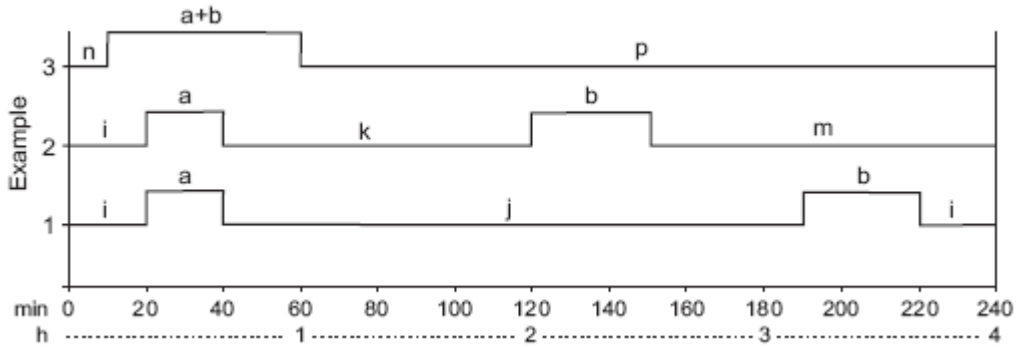

Figure 3.6. An example of the feeding behaviour structure for three individual animals throughout a 4-hour period. Raised values of the signal denote periods of feeding, whereas lowered values denote inter-meal intervals. The length of feeding duration and inter-meal intervals are represented by letters. Note: 2i+j = i+k+m = n+p

Table 3. 4. Calculated feeding behaviour parameters for each example shown in Figure 3.6.

Example	Average duration	Total duration	Minimum duration	Maximum duration	Number of meals	Average inter-meal interval	Total inter-meal interval	Minimum inter-meal interval	Maximum inter-meal interval	Number of inter-meal intervals
1	(a+b)/2	a+b	a	b	2	(2i+j)/3	2i+j	i	j	3
2	(a+b)/2	a+b	a	b	2	(2i+j)/3	2i+j	i	m	3
3	(a+b)	a+b	a+b	a+b	1	(2i+j)/2	2i+j	n	p	2

3.2.6. Data Cleaning

As error could be introduced into the dataset due to any hardware malfunction in the GrowSafe™ system, the summarized data needed to be cleaned before further processing could be considered.

3.2.6.1.Sources of Data Error

Within a raw dataset the possibility of distinguishing the difference between system malfunctions and true animal absence (both were recoded as 0) was limited. Therefore, the first step in data cleaning involved removal of 0 values during those periods of time when the system was not functioning. System failure occurred most frequently in individual panels. Hardware configuration in this study was such that four panels covered five pens; consequently the failure of one panel affected the data quality of more than one pen (see Figure 3.1). The information collected by each sync chip was used to identify when system failures (by panel) had occurred. A program was written in Visual Basic 6.0 that combined processed sync chip and animal feeding behaviour data. This resulted in the identification of missing values in place of 0 when the system was not working properly. In cases where a panel spanning two distinct pens failed, data collected for both pens (even if neighbouring panels were functioning properly) were affected, and data for all affected pens were set to missing.

3.2.6.2. Determining thresholds for data use based on system performance

Under ideal conditions (100 % performance) the GrowSafe™ hardware used in this study would record the presence of each animal at the feed bunk every 5.25 s.

However, the collection system was configured to record only integer time values, thus an ideal scanning/recording rate of 6s was assumed. Given this read rate, a total of 2400 sync chip readings (4 h* 3600s per h per 6 s) would be expected in a 4 h period. However, it was unrealistic to expect such performance with any RFID technology and therefore it was important to define reasonable limits for the exclusion of poor quality data based on system performance. Two factors were considered when evaluating system performance, and through this, data quality: system read rate and the length of time between sync chip readings. For each data collection period, the number of sync chip readings and the maximum length of time between two consecutive readings were recorded; this information was later used to define data quality thresholds. These thresholds were determined using a read rate rule requiring the maximum length of time between 2 consecutive readings to be less than 300 s. This particular length of time was selected with the definition of 'meal' in mind, in which case if the system was not functioning for less than 300 s the data would still be valid. However, if the system was malfunctioning for less than 300 s but more frequently within a period, the accumulation of faulty periods would yield inaccurate predictions of feeding behaviour. The incidence of system malfunction was defined in terms of percent sync chip availability. This was done by dividing the actual number of sync chip readings by the expected number of readings and multiplying by 100. Thus, for each individual panel, a value indicating the percent of data to be removed was calculated for selected percent availability values as illustrated in Figure 3.7.

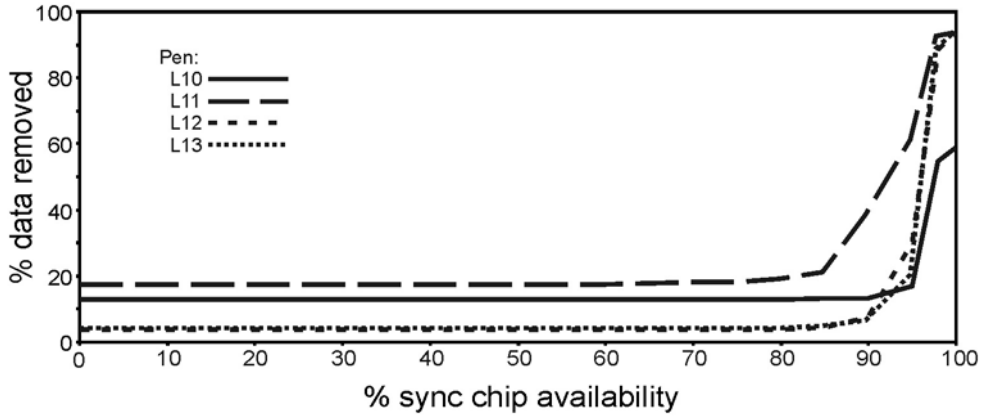

Figure 3.7. GrowSafe™ panel functionality based on sync chip performance. The data not meeting the criteria of 2400 readings per 4 h period increases exponentially starting at 85 % sync chip availability.

Using Figure 3.7 as a guide, it was concluded that, for data to be acceptable, a 4-hour period should contain a minimum of 2040 sync chip data readings (85 % availability), given that at this point the percent of data removed increased exponentially in each pen. The maximum length of time between 2 consecutive readings was set to < 300 s. This was summarized in the following formula:

For each $x_2 \in \{0, 1, 2, \ldots, 100 \mid x_2 \in \mathbf{N}\}$:

$$F(x_1,x_2)= \begin{cases} \text{Recorded value} & \text{if } x_1<300 \\ \text{“.”} & \text{otherwise} \end{cases} \quad (3.1)$$

where x_1 is the time interval in seconds between two consecutive system scans and x_2 is the minimum percent accuracy required for system robustness. If the defined data quality requirements for a specific period were not met, the data for that period were set to missing. Following the completion of this step the dataset was considered clean and acceptable for input into the data reduction routine.

3.2.7. Data Mining

Data mining, in general, deals with the discovery of non-trivial, hidden and interesting knowledge from different types of data.

3.2.7.1. Dataset Reduction

The dataset was reduced such that only feeding behaviour data 10 d prior to a M animal being removed from its home pen were used for further analysis. This decision was based on the fact that most animals were removed from the pen within the first 10 d of the trial. Extending this time period beyond 10 d would have resulted in a very small dataset, simply because of the lack of morbidity in the cattle population. A program was written in Visual Basic 6.0 to create a dataset based on this 10-d rule. The dataset was further reduced to include data from a sample population containing all M cattle and a matching number of H cattle. In this manner, a 1: 1 M to H ratio of animals was obtained from the original dataset such that for every M animal an H animal was selected from the same pen, on the same day. This approach ensured that subjects were selected under similar environmental and feedlot management conditions. This resulted in the 10 d of data captured in the dataset being unique for each healthy-sick pair in that the starting point was defined as the day prior to the first day M cattle were removed from their pen for medical assessment and treatment. The dataset was then constructed using information collected over 10 d prior to the point that M cattle were identified and removed from the pen. For incidences where the animal was first removed from the pen within the first 10 d of the experiment, only data from the beginning of the experiment to one day prior to removal of the animal from the pen were selected for that particular healthy-sick pair.

The input parameters to the Visual Basic 6.0 program for creating the dataset were a set of M and its pair-wise contemporary healthy animal transponder ID tags, the removal dates of all the M animals, and the number of days selected for analysis prior to the animal's removal from the pen. The number of days selected for analysis varied from one day to up to 10 days, depending on when cattle were removed for morbidity. If they were removed prior to spending a minimum of 10 days at the feedlot, data were collected only for the number of days that the animals were there. The algorithm extracted information for the specified day for all M and H animals. Given the CLS classification definitions, our confidence that the M group that fell in the low category was not strong. For model development purposes we wanted to only include M animal data for which we had a high degree of confidence. Therefore, only data for animals with high and moderate CLS categories were included. Furthermore, to reduce data quantity and in an attempt to increase the accuracy of the final model, only the periods of the day in which the animals were most active at the feed bunk were used for subsequent analysis. Active feeding periods were determined by plotting the diurnal feeding behaviour of 10 H and 10 M animals over a 5 d period prior to being removed from their home pens. Figure 3.4 illustrates that peak feed bunk activity occurred between the hours of 0600 and1800, and therefore only Periods 2 (0600-1000), 3 (1000-1400), and 4 (1400-1800) for each of the 10 d prior to removal from their home pen for medical assessment were included in the dataset, resulting in a total of 30 periods per animal (Figure 3.8). The number of periods was less for cattle that were removed from their pen within the first 10 d at the feedlot. This reduced dataset was normalized using Proc STDIZE in SAS (1991) to reduce any skewing caused by large variances in the data.

Figure 3.8. Highlighted periods represent periods included in the dataset.

3.2.7.2. Principal Component Analysis (PCA)

Due to the high dimensionality (12 variables: feeding duration, inter-meal interval (min., max., avg., SD and total; min/d), feeding frequency (visits/d) and number of inter-meal intervals) of the dataset, Principal Component Analysis (PCA) was employed to condense the data into fewer dimensions without excessive loss of information. The application of PCA to the data resulted in capturing most of the variability within a dataset. This allowed for the comparison of feeding patterns between animals as well as changes within an animal across the 30 time periods assessed. The first five PCs identified in the dataset cumulatively captured more than 99 % of the variability in the dataset. Based on these results, the first five PCs were selected to construct a revised dataset that was later used as input data for the clustering procedure.

3.2.7.3. Clustering

The clustering technique used in this study was performed by the FASTCLUS procedure in SAS (1991).

One of the options of the FASTCLUS procedure allows the user to indicate the number of clusters the algorithm should divide the objects of the dataset into. Eight clustering strategies (setting the number of clusters) were examined including 2 to 9 clusters with each consecutive run of the clustering algorithm. This resulted in 8 separate

and distinct output datasets for further analysis. Upon close examination of clustering outputs and cluster membership results, it became evident that the upper and lower limits of the number of cluster strategies to be used for further calculations needed to be defined. The average number of animals in a cluster (using a cluster strategy of 9 clusters) was low, and therefore not a good representation of H or M animal feeding behaviour. Given that a minimum of two groups (i.e. H and M) were expected to emerge from the clustering, the lower limit was naturally defined as two. The upper limit was selected by consequently testing the performance of the algorithm using each cluster strategy, to the point where overall model performance started to decline. In this case, 6 clusters. Consequently, the number of clusters considered changed between 2 and 6, inclusive.

3.2.7.4. Classification

Classification is defined as a task where data points are assigned to predefined classes. In other words, classification requires supervised learning, where the input data must specify what is to be learned, whereas clustering is an unsupervised task, and thus the clusters are not specified in advance. In this study, for each n-cluster dataset (where n Є {2, 3, 4, 5, 6}), cluster membership was examined. Based on the percentage of M animals belonging to each individual cluster, clusters were labelled as morbid-clusters (M-cluster) or healthy-clusters (H-cluster). Three thresholds for morbid cluster designation were set: 45, 50 and 55 % M membership clusters, where the number of animals classified as morbid corresponds to percent classification for each of the clusters. Clusters that were not M-clusters were defined as H-clusters. The input data were analysed repeatedly using each definition, resulting in a total of 15 unique output

datasets, including: n-cluster-45, n-cluster-50 and n-cluster-55 datasets (where n Є {2, 3, 4, 5, 6}). Initially, the 50 % threshold was selected as the central definition on the rationale that if more than half of animals (i.e. >50 %) were morbid, then it must be an M cluster. The examples described from this point forward will use a membership distinction of 50 %. To access the soundness of the selected 50 % threshold, cluster memberships of 45 % and 55 % M membership were also tested. Based on these parameters, all animals in each cluster were assumed to inherit the apparent health status designation of that particular cluster. In other words, if a cluster was defined as an M-cluster, then all animals belonging to that cluster were assumed to have an apparent status of M for that time period. (Figure 3.9 and Table 3.5)

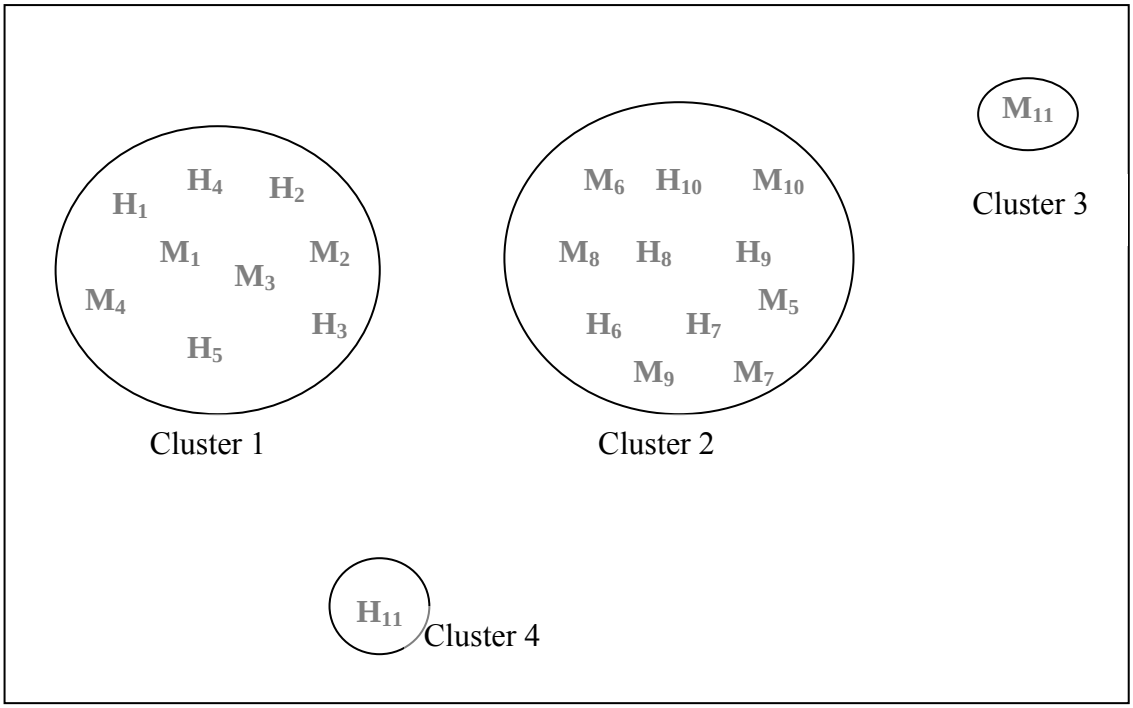

Figure 3.9. A 4 cluster example demonstrating the distribution of healthy (H) and morbid (M) animals within each cluster. Cluster designation will differ, depending on the threshold used (45, 50 or 55 %).

Table 3. 5. Apparent status cluster classification given the example in Figure 3.9.

Cluster	% morbid	Animal	Status	Apparent status of cluster		
				45 % threshold	50 % threshold	55 % threshold
1	44	H1-5	H	M	H	H
		M1-4	M	M	H	H
2	55	H6-10	H	M	M	H
		M4-10	M	M	M	H
3	100	M11	M	M	M	M
4	0	H11	H	H	H	H

In the example shown in Figure 3.9, 11 M and 11 H animals were clustered into 4 non-overlapping groups. In Figure 3.9 clusters 3 and 4 contain only one element each, suggesting that these two datapoints may be outliers. This example also demonstrates why choosing the right number of clusters is important, and that clustering can be used for outlier detection. Outliers may emerge as single data points or as small clusters far removed from the main clusters. To do outlier detection at the same time as clustering the entire dataset, the sufficient use of clusters is important to represent both the main dataset and the outliers. As indicated in Table 3.5, 44, 55, 100 and 0 % of the M cattle belonged to clusters 1, 2, 3 and 4, respectively. Furthermore, at a 45 % threshold level definition, clusters 1, 2 and 3 were defined as M and only cluster 4 was defined as H. Therefore, all member(s) of clusters 1, 2 and 3 were given the apparent status of M, whereas the member in cluster 4 was given the apparent health status of H. Similarly, at a 50 % threshold level, clusters 1 and 4 were defined as H, whereas clusters 2 and 3 were defined as M. Thus all member(s) of clusters 1 and 4 were given the apparent status of H for that time period, whereas member(s) of clusters 2 and 3 were assigned an apparent status of M. The same rules were applied to the 55 % threshold level definition.

3.2.8. Pattern Recognition

The pattern recognition process consisted of two major steps, the first being the creation of a string of length 30, i.e. a string comprising of 30 elements, each element representing a 4 h period, later referred to as a binary string. The second step defined a 'time window algorithm'.

3.2.8.1. Creation of a Binary String

In this experiment, a binary string (B) was defined as an arbitrary sequence of H's and M's that could be transposed into an array of 0's and 1's by assigning H a value of 1 and M a value of 0. The rationale behind the creation of this binary string was to develop a method of quantifying feeding behaviour for each 4 h period where data were observed for each M and H pair. A description of how the binary string was created is as follows. Each animal had feeding behaviour data that were summarized into PCs for a 10 d (or less) period of time. The 10 d sample was broken into 30 4-h-periods (3 periods/d over 10 d). For future reference, let this set of data be referred to as "The dataset". The dataset was then used as input for the K-means clustering algorithm. This gave rise to 5 new and distinct clustering strategies (2,3,4,5 and 6 number of clusters), and with those 5 new output datasets evolved; dataset-n-cluster, where n Є {2,3,4,5,6}. Furthermore, each of these resulting datasets were subject to 3 definitions of cluster membership: 45, 50 and 55 % M membership clusters, resulting in 15 datasets D, such that D={dataset-n-cluster-m | n Є {2,3,4,5,6} and m Є {45 %, 50 %, 55 % M membership}. Let d Є D (i.e. any given dataset from D). For all d Є D, d contained an apparent health status classification for each selected time period and each M animal and its contemporary H

pair. Thus, selection of 3 cluster membership possibilities and 5 cluster strategies resulted in 15 apparent health status classifications for each animal. Each apparent health status, concatenated over the time period of 10 d produced a binary string with a length of 30 for each animal. The first position was identified as period 1, then consecutively, the second as period 2 and so on. Period 1 was defined as the period immediately prior to the animal being removed from their home pen, and period 30 being 10 d prior to the animal being removed from its home pen (Figure 3.10.).

To better visualize and analyze the apparent state of each animal over a 10 d period, the median apparent health status in pre-defined sliding windows of consecutive periods was calculated for each animal. A sliding window is a dynamic string, containing a subset of a binary string. The different window sizes considered were: 3, 5, 7, 9, 11 and 13 4-h periods; with W set as the set of all window sizes. Window sizes were selected based on the rationale that a minimum of 1 day's data (i.e. 3 periods) were required to be able to make a decision. Window sizes of odd length were examined in order to avoid a tie between H and M declared health statuses. Each one of the 15 strings was then examined by one of these moving windows to return a declared health status value for each animal under each combination of cluster strategy and threshold level. An upper limit value of c was determined for each window size as follows:

$$c=\text{ceiling}((w+1)/2) \quad (3.2)$$

where w Є W, to compare with *a*, where *a* is the # of M apparent status classifications within that particular window. (Note: the ceiling function returns the closest integer that is greater or higher than the input value.) For each animal and each window size, if any of the sliding windows returned an *a* value greater than c, then the animal was declared M; otherwise the animal was designated to have H status. In other words, the method

uses a sliding window technique to control the length of the period to be matched against c. Assume that b=[b_1, …, b_w], where b represents the span of the initial window covering the first w integers of the binary string B. The sum of all apparent M status classifications within w was calculated and compared to c. If the sum of all apparent M status classifications exceeded c, then the animal was declared M, otherwise the window would slide one position to the right, leaving b2, …, bw+1 for rule matching. This process was continued until the end of the string was reached. If the animal was not classified as M throughout the process, then it was assumed to be H.

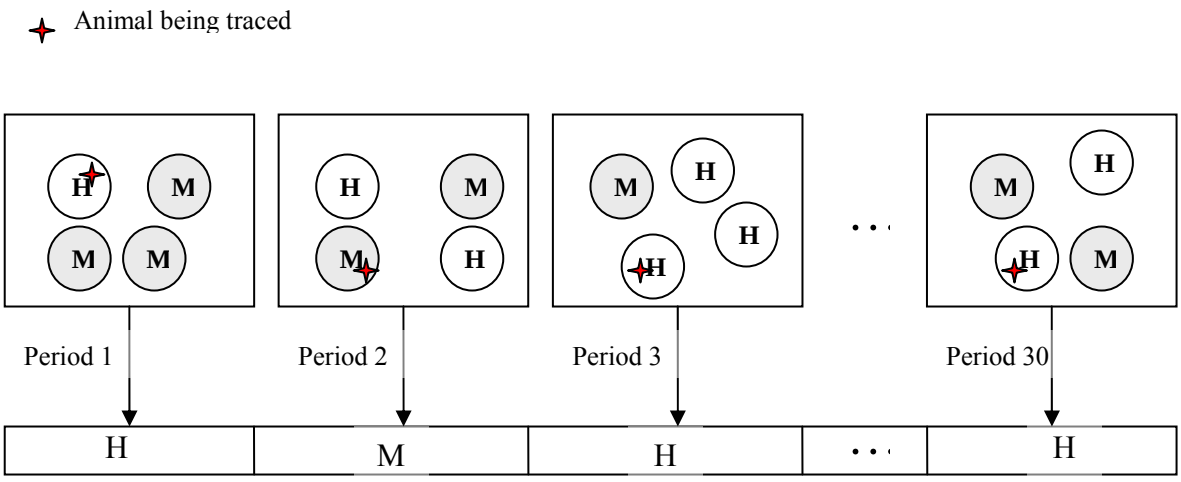

Figure 3.10. Each period of the graph represents a 4 cluster example where each cluster is labelled with an apparent status as defined by a 50 % threshold level definition. The animal being traced is shown to inherit the apparent status of the cluster it belongs to in each period, creating the binary string.

3.2.9. Defining and Selecting a Model

An optimal 'model algorithm' was defined as a ratio between actual health status, as indicated by whether the animal had been removed from it's pen for sickness by the pen checker, versus the declared health status definitions. Three scenarios were used to define the best model and include:

1. 100 % H model: 100 % H accuracy and highest percent M accuracy. Animals were classified into two groups, one that contained only H animals, the other group including the rest of the animals. The animals were assigned to each respective group such that the percent M accuracy in the M group was maximized, without jeopardizing the 100 % accuracy of the H group.
2. 100 % M model: 100 % M accuracy and highest percent H accuracy. Animals were classified into two groups, one that only contained M animals, the other group including the rest of the animals. The animals were assigned to each respective group such that the percent H accuracy in the H group was maximized, without jeopardizing the 100 % accuracy of the M group.
3. Overall model: Highest percent of M and highest percent of H accuracies. Animals were classified into two groups such that the percent accuracies of M and H animals were maximized in each group.

It is important to note that 100 % H accuracy does not necessarily mean that no M animals were classified into that group. A 100 % H classification would include all healthy animals, and perhaps other M animals that behaved like H animals. However, within that same model, the M group would only include animals that have been classified as M since all healthy animals were members of the H cluster. Therefore, by

setting the standards of H to 100 % accuracy only a portion of the M animals would be classified correctly.

3.2.10. Creating a Naive Dataset

A common scenario in creating data mining models is to predict their accuracy by comparing them against a naive dataset. This prevents the problem of over-fitting (making the model too specific for the model dataset (Goodner *et al.*, 2001), and gives a better measure of the accuracy of the generated models.

3.2.10.1. Description of the Naive Dataset

Three hundred and eighty-four mixed breed British x Continental feedlot steers, averaging 322 ± 34.7 kg, initial BW were monitored over a 142 d feeding period in the same four feedlot pens at the Cactus Feeders feed yard in Ararillo, TX as previously described for the modelling dataset. Cattle were received at the study site from sources in Kansas, Oklahoma and Nebraska between February 11 and February 14, 1998. From receipt until allotment, steers were maintained in holding pens and fed a standard receiving ration consisting of a moderate concentrate mixed diet plus loose, long-stem alfalfa hay and allowed free access to drinking water. Upon arrival, cattle were processed by administration of an IBR – Leptospira modified live vaccine (Vista 5 L5 SQ, Intervet Animal Health Inc.); a 7-way clostridial bacterin-toxoid (Vision-7®, Intervet Animal Health Inc.); a drench containing 1,000,000 IU vitamin A and 200,000 IU vitamin D (Rovimix dispersible liquid, Roche Vitamins Inc.) and treated for parasites (Dectomax®, Pfizer Inc.). Animals were re-implanted on April 5, 1998. Cattle were fed three step-up diets containing 36, 29 and 18 % roughage (DM basis). Diet transitions were made over two days, with the lower energy diet fed at the first two feeding cycles

on day 1 and higher energy diet fed on the last two feeding cycles on day 2. Cattle were fully transitioned to the 10 % roughage finishing diet by March 7, 1998 (Table 3.6.). The three transition diets were fed three times daily at approximately 0600, 1030 and 1300 hours. The finishing diet was fed twice per day at 0600 and 1300 daily. All diets were formulated to meet or exceed National Research Council (1996) requirements for growing – finishing beef cattle. Feed bunks were visually evaluated and scored for the amount of residual feed at approximately 0600 hours daily. Cattle were fed to appetite, with the amount of feed issued to each pen adjusted daily by the amount of feed, if any, remaining in the bunk prior to the first feeding of the day. Bunks were managed throughout the finishing period to minimize the amount of residual feed carried over from day to day. Lung lesion and liver abscess information was collected on all animals at the time of slaughter.

Table 3. 6. Ingredient and nutrient composition of transition and finishing diets used in Naive Dataset

	Transition Diets			Finishing Diets	
Item	Ration 1	Ration 2	Ration 3	LP /	R / T
First date fed	02/17/98	02/22/98	02/28/98	03/06/98	03/06/98
Last date fed	02/21/98	02/27/98	03/05/98	07/07/98	07/07/98
Total days fed	5	6	7	123	123
Ingredient, %					
Steam-flaked corn	47.70	56.30	50.90	53.8	53.8
High moisture corn	0.0	0.0	14.60	19.5	19.5
Alfalfa hay, chopped	15.70	15.90	17.20	8.1	8.1
Cottonseed hulls	19.90	13.40	0.00	0.00	0.00
Corn silage,	0.00	0.00	4.20	4.2	4.2
Molasses	7.10	5.40	4.00	2.0	2.0
Animal fat	0.00	0.0	2.00	4.1	4.1
Starter Supplement	9.10	9.00	0.00	0.0	0.0
Finisher Supplement	0.00	0.00	7.00	8.3.	8.3.
Micro-ingredients	0.50	0.00	0.00	0.00	0.00
Calculated Composition					
Dry matter, %	81.2	79.4	72.7	72.7	72.7
$NE_{m,}$ Mcal / 100 lb	77.9	83.6	94.0	100.7	100.7
$NE_{g,}$ Mcal / 100 lb	49.9	55.0	64.1	69.8	69.8
Crude protein	13.7	13.9	13.8	13.8	13.75
Non-protein N, %	1.80	1.80	2.70	3.17	3.17
Crude fat, %	2.60	2.80	5.0	7.23	7.23
NDF, %	29.2	24.7	16.6	12.81	12.81
Calcium, %	0.67	0.64	0.72	0.66	0.66
Phosphorus, %	0.33	0.34	0.31	0.32	0.32
Potassium, %	1.25	1.12	1.07	0.79	0.79
Magnesium, %	0.25	0.23	0.25	0.24	0.24
Vitamin A, IU/lb	3,331	3,000	1,964	1,784	1,784
Vitamin D, IU/lb	333	300	196	178	178
Vitamin E, IU / lb	10.0	9.0	0.0	0.0	0.0
Aureomycin, g/ton [a]	931	42.0	39.3	35.7	35.7
Cattlyst, g/ton [a]	0.00	11.1	11.1	11.1	11.1
Rumensin, g/ton [a]	0.00	20.0	24.0	--	27.8
Tylan, g/ton [a]	0.00	11.0	11.0	--	9.0

[a] Hand-weighed and added to the conventional rations via water slurry.

3.2.11. Applying the Model Algorithm to the Naive Dataset

Based on CLS classification, a subset of thirteen M:H pairs were identified from the naive dataset. Raw feeding behaviour data were summarized into behaviour data

variables as described in Section 3.2.4. The algorithm described in Sections 3.2.5 and 3.2.6 were used to clean the data. Dataset reduction, data normalization and PCA routines (Sections 3.2.7.1, 3.2.7.2) were also applied resulting in a dataset consisting of animal IDs and the first 5 PCs derived by PCA analysis for each of the 30 4-h processing periods. The Euclidean Distance Formula (Duda *et al.*, 2001) was used as previously described to assign animals to one of the pre-defined clusters at which point the animal inherited the apparent health status of that cluster. Consequently, each animal was given an apparent health status for each one of the 30 processing periods as described in the Classification Section (3.2.7.4). The method of creating the binary string, as described in Section 3.2.8, was used to create a binary string representing animal health status, the sliding window technique was implemented to state the declared health status of each animal from the naive dataset. Declared results were then compared with the actual health status of the animal. Results of the comparison were stated in terms of percent accuracies.

3.2.12. Descriptive Statistics

Descriptive statistical methods were applied at two points within this study. First, the Mixed Liner Models Procedure (SAS, 1991) was used to calculate least squares means of a 2X2 factorial design of animals removed or not removed from their home pens by pen checkers, with or without lung lesions at slaughter. The Means Procedure (SAS, 1991) was also implemented to derive simple statistical information such as the percentage of animals with lung lesion that were never removed from their home pens by the pen checker. Finally, differences and similarities between the model and naive

Datasets were defined by calculating correlation coefficients between all variables present in each dataset using the Correlation Procedure (SAS, 1991).

3.3. Results

3.3.1. Animal Data and Descriptive Statistics

Out of the 384 animals used for the model dataset, 16 animals were rejected from the study and were sent prematurely for slaughter, 9 animals died, 93 were removed by the pen checker, and the remainder (n=267) were classified as 'Other' as they did not fit into the 3 categories previously described (Table 3.7).

Table 3. 7. Summary of the number of animals falling into removed (animals that have been removed from their home pen for medical assessment), dead, reject or other categories within the model dataset.

	Total number of animals (n=384)					
	Reject (n=16)		Dead (n=9)		Removed (n=93)	Other [a] (n=267)
	Removed (n=15)	not Removed (n=1)	Removed (n=3)	not Removed (n=6)		
Lung Lesions	1	0	0	0	17	28
no Lung Lesions	14	1	3	6	76	239

[a] The category 'Other' includes all healthy animals as well as animals with liver abscesses that were never removed from their home pens for morbidity by a pen checker for treatment.

Table 3.8 shows how the number of animals related to the number of days an animal had been on feed before it was removed by a pen checker for exhibiting signs of morbidity for the first time. The largest noteworthy difference between the model and

naive Datasets is the total percentage of M animals in each CLS category. In the model dataset, 75 % of M animals were categorized as having moderate or high CLS, whereas in the naive dataset only 33 % of M animals fell into either of these categories. Animals that were removed from their home pens for the first time for morbidity by the pen checker within the first 14 d on feed accounted for 75 % and 83 % of the model and naive Datasets, respectively. From this group of animals, only 28 % were categorized as having low confidence level of sickness in the model dataset, compared to 70 % in the naive dataset.

Table 3. 8. Percentage of the total number (n) of animals assigned to the high (Hi), moderate (Mo) and low (Lo) Confidence Level of Sickness categories in both the model and naive Datasets.

Days on Feed	Model Dataset (n=93)		Naive Dataset (n=53)	
	% Removed as Morbid	CLS categories	% Removed as Morbid	CLS categories
1 to 14	75 % (n=70)	Lo: 28 % (n=18) Mo: 46 % (n=32) Hi: 26 % (n=20)	83 % (n=44)	Lo: 70 % (n=31) Mo: 16 % (n=7) Hi: 14 % (n=6)
15 to 28	10 % (n=9)	Lo: 12 % (n=1) Mo: 44 % (n=4) Hi: 44 % (n=4)	4 % (n=2)	Lo: 0 %(n=0) Mo: 100 % (n=2) Hi: 0 % (n=0)
29 +	15 % (n=14)	Lo: 29 % (n=4) Mo: 64 % (n=9) Hi: 7 % (n=1)	13 % (n=7)	Lo: 57 % (n=4) Mo: 43 % (n=3) Hi: 0 % (n=0)

Differences in the feeding behaviour of two distinct experimental groups are shown in Table 3.9. All feeding behaviour variables with the exception of bunk attendance frequency and maximum inter-meal interval were higher ($P<0.005$) in the naive dataset than the model dataset with the exception of maximum inter-meal interval and bunk attendance frequency which were greater in the model dataset.

Table 3. 9. Comparison of feeding behaviour variable (mean ± SE) summaries between the model and naive Datasets summarized by 4-hour periods.

Variable	Model Dataset	Naive Dataset
Average meal duration (min)	9.39 ± 0.02^{A}	9.74 ± 0.02^{B}
Total meal duration (min)	16.74 ± 0.03^{A}	16.85 ± 0.03^{B}
Minimum meal duration (min)	7.07 ± 0.02^{A}	7.48 ± 0.02^{B}
Maximum meal duration (min)	12.10 ± 0.02^{A}	12.38 ± 0.02^{B}
Bunk attendance (visits)	1.28 ± 0.002^{A}	1.22 ± 0.002^{B}
Average inter-meal interval (min)	147.19 ± 0.10^{A}	148.38 ± 0.14^{B}
Minimum inter-meal interval (min)	115.67 ± 0.14^{A}	117.88 ± 0.19^{B}
Maximum inter-meal interval (min)	185.53 ± 0.07^{A}	184.87 ± 0.09^{B}

^{A,B} within a row, values followed by different letters differ ($P<0.005$).

3.3.2. Clustering

The choice of number of clusters is an important sub-problem of clustering. Figure 3.11 demonstrates the percent accuracies of each of the three models. The highest 100% M model performance accuracy was 58 % in a 4 cluster situation. The 100 % M model performed the best when the 5 cluster strategy was used, reaching an accuracy of 68 %. This accuracy declined to approximately 55 % when the 6-cluster strategy was applied. The overall model performed comparably well through cluster strategies 3 to 6. The percent accuracy of the overall model increased 5 % between cluster strategies 3 to 5, but a reduction of 6 % accuracy was observed between the 5^{th} and 6^{th} cluster strategies.

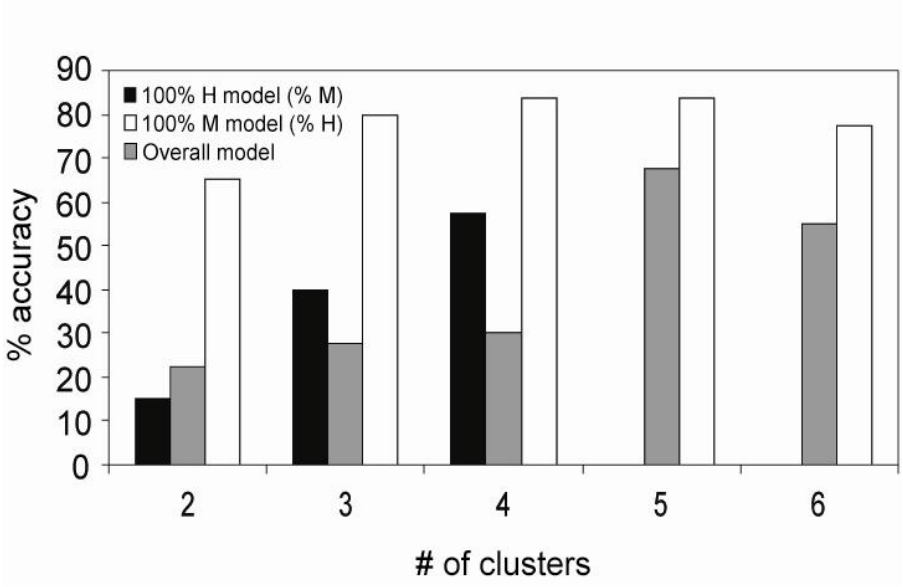

Figure 3.11. Ultimate classification accuracies of the three models (100% H, 100% M, and Overall models) using 2 to 6 cluster strategies on the model dataset.

As shown in Figure 3.11 the percent accuracies of all three models started to decline when the input data were divided into more than 5 clusters. As a result, the pattern recognition process was not applied to cluster strategies with cluster numbers greater than 6.

3.3.3. Models

A total of 126 models (6 cluster sizes, 3 thresholds and 7 window sizes) were applied to the model dataset. Figure 3.12 shows the performances of all models in terms of percent H and percent M accuracies.

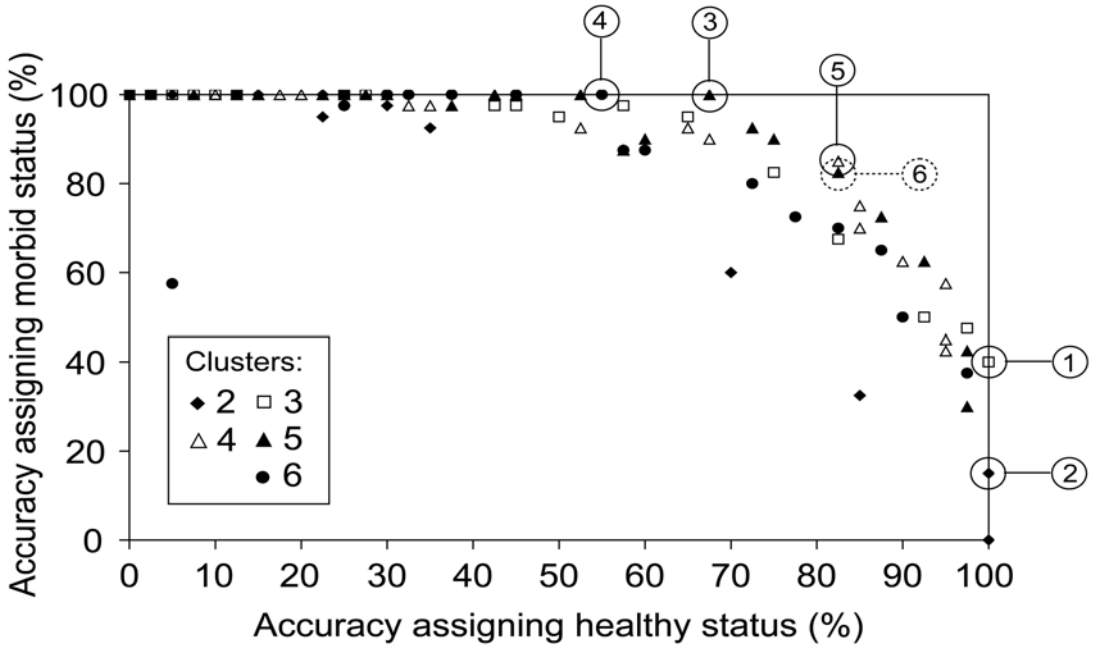

Figure 3.12. Model dataset results. Healthy and Morbid percent accuracies are indicated by each data point representing each unique (combination of number of clusters, cluster classification threshold levels and window size) classification model. Numbers 1 and 2 indicate the top two 100 % H models, numbers 3 and 4 indicate the top two 100 % M models and numbers 5 and 6 highlight the top two overall models.

As per model definitions described in this study in Section 3.2.9, the particulars of each top two models derived from the model dataset are summarized in Table 3.10. The optimal 100 % H model (indicated as # 1 in Figure 3.12) predicted 40 % of the M animals. These animals were categorised as M between 1 to 5 d (on average of 3.7 d) earlier than visual observation by a pen checker (Table 3.10). The second optimal 100 % H model (#2) only predicted 15 % of the sick animals correctly, but up to 7 d, and on average 4.7 d earlier than the time when the pen checker removed the animals for sickness (Table 3.10). In contrast, the top two 100 % M models (#3) predicted H comparably. The algorithms were able to predict the M animals up to 6, (average of 3.5) d earlier (Table 3.10), and with 67.5 % accuracy with the 100 % M model , and 7, (average of 4.5) d earlier (Table 3.10) and with an accuracy of 55 % with the second 100 % M model (#4). Both of the top two overall models (#s 5and 6) predicted H animals

with a 82.5 % accuracy, and M with 85 and 83 % accuracies, with an average of 3.3 and 5 d earlier than a pen checker, respectively (Table 3.10).

Table 3. 10. Model summaries for morbid and healthy cattle as well as average early prediction number of days within the model and naive Datasets

Model type		Cluster size	Threshold level	Window size	Model Dataset					Naive Dataset				
							Predicted early (d)					Predicted early (d)		
					%H	%M	Min	Mean	Max	%H	%M	Min	Mean	Max
100% H	1	3	55 %	15	100	40	1	3.7	5	100	0	n/a	n/a	n/a
	2	2	55 %	9	100	15	0	4.7	7	100	0	n/a	n/a	n/a
100 % M	3	5	50 %	11	67.5	100	1	3.5	6	58.3	67	2	1.6	6
	4	6	50 %	9	55	100	1	4.5	7	25	75	1	2.25	6
Overall	5	4	50 %	11	82.5	85	0	3.3	6	100	50	1	1.2	2
	6	5	55 %	7	82.5	83	0	5	7	83.3	58.3	1	1.4	6

3.3.4. Naive Dataset

The model results derived from Figure 3.12 were compared to results obtained by applying the model algorithm to the naive dataset (Table 3.10). None of the 100 % H models were able to predict morbidity with all animals being predicted as H. However, the best 100 %M (#3) (Figure 3.13) model predicted 58.3 % of the H and 67 % of the M animals on average 1.6, and up to 6 d earlier than traditional methods (Table 3.10). The second 100 % M model (#4) (Figure 3.13) predicted 25 % of the H and 75 % of the M correctly, on average 2.25, and up to 6 d earlier than the pen checker (Table 3.10). The overall model (#5) predicted the H animals with 100 % accuracy, whereas the M cattle were only predicted with 50 % accuracy (Figure 3.13). When model (#6) was mapped onto the naive dataset (i.e. the set of rules and procedures developed using the model dataset were applied to the naive dataset), the result was a prediction of 83.3 % accuracy for H cattle and 58.3 % accuracy for M cattle (Figure 3.13). Even though H and M animals were represented equally in the naive dataset, it is unknown why H was still predicted more accurately in both overall models. One possibility may be that the parameters of the algorithm developed in this study were set such that it allowed for more H animals to be classified as M.

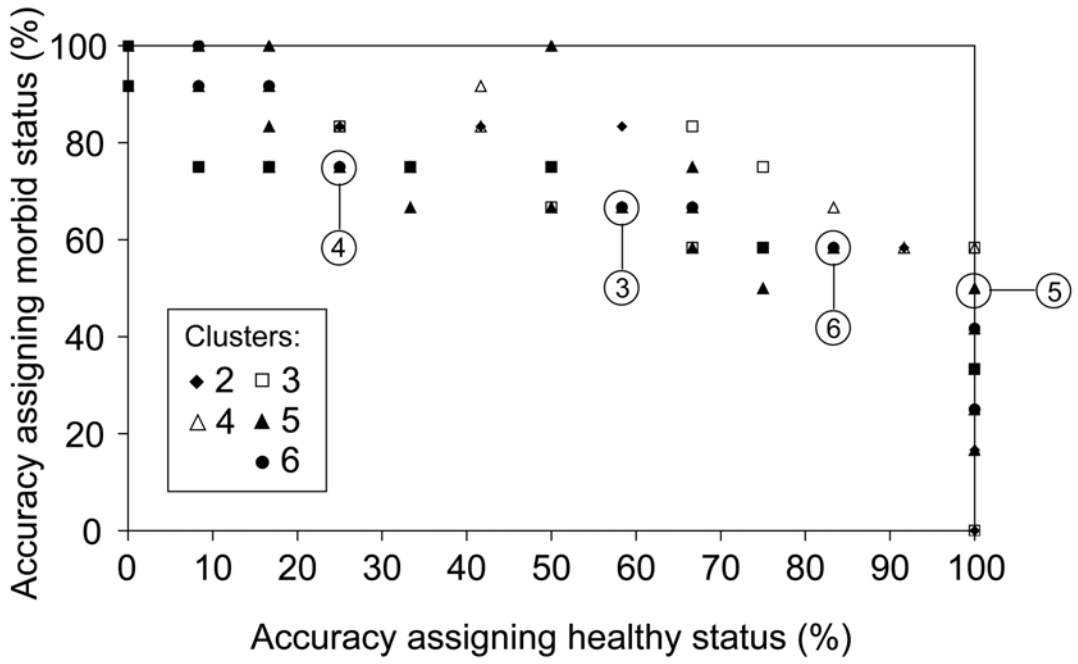

Figure 3.13. Naive dataset model results. Healthy and Morbid percent accuracies of each unique model after each individual animal from the naive dataset has been classified using the classification algorithm derived using the model dataset. Numbers 3 – 6 indicate the accuracies at which the best performing models highlighted in Figure 3.12 performed using the naive dataset.

3.4. Discussion

3.4.1. The Datasets

To support the theory that current methods of sickness detection are not optimal, findings derived in this study (Table 3.5) indicate that 63 % of the animals with lung lesions in the model dataset were never identified as being removed from the pen for illness. This is not necessarily due to pen checker error as it is possible for animals to develop lesions without exhibiting clinical symptoms, or the lesions may have formed in the lungs prior to the arrival of the cattle at the feedlot. Contrary to this, the concept that most animals become sick within a short period after arrival to the feedlot (Griffin, 1998; Mathison, 1993; Smith, 1998) is strongly supported by data in both the model and naive Datasets.

Differences in feeding behaviour were observed between the 2 trials used in this study. These differences indicate that even under similar feedlot management conditions, feeding behaviour between groups of cattle can vary. These differences may be attributed to several factors including weather, source of cattle, length of transport to the feedlot, type of feed, sex, animal interactions, etc.

3.4.2. Modelling Strategy

The goal of this study was to develop an algorithm that could identify patterns of morbid feeding behaviour prior to being detected by a pen rider. One of the key steps in developing this model was the categorization of animals into CLS categories. Use of only the moderate and high CLS categories increased confidence levels to indicate that the modeling of "truly sick" and healthy animals was captured. The modeling strategy used in this study allowed the algorithm to deal with 'normal' fluctuations of feeding behaviour. This was achieved by setting certain boundaries and threshold levels in these boundaries and then testing them. The idea was to allow M animals to behave as H and H animals to behave as M animals a fraction of the time, on the fundamental assumption that healthy cattle would at times have feeding behaviour patterns that were similar to M cattle. Ultimately, it is important that the algorithm be robust enough to be used on any feeding behaviour dataset collected at a commercial feed yard. Consequently, the algorithm should be useful in predicting M animals sooner at a multitude of different feedlot locations, housing a variety of different breeds and applying numerous management strategies under varying environmental conditions across years. Therefore, an important part of the modeling strategy was to test the algorithm developed on the

model dataset with a naive dataset. Each specific output of the model will be discussed in more detail in the following sections.

3.4.2.1.Number of Clusters

Olofsson (1999) found that cattle classified as "dominant" increased their feeding duration when bunk space was reduced from 90 cm to 23cm per animal, whereas animals that were classified as 'subordinate' altered their intake pattern as well as fed more often during the less preferred hours of the day. In a study conducted by Hickman *et al.* (2002) it was concluded that cattle that exhibited the highest average daily gain and were the most feed efficient, also had the greatest variation in daily feeding patterns. The above findings support that many factors contribute to variation in feeding behaviour that may result in grouping or clustering individuals based on several factors concurrently. Schwartzkopf-Genswein *et al.* (2003) stated that cattle feeding behaviours are inherent and not easily altered. This suggests that even though the onset of morbidity in cattle may alter feeding behaviour, innate behaviours such as reaction to environmental factors may still dominate or override these more suble behaviours, causing animals to cluster into two or more groups, irrespective of their health status. Defining the number of such clusters and which variables would be the driving force in defining these clusters was a challenge not only because the data were variable, but also because of the nature of the data and the clustering method used. Dy and Brodley (2004) also describe the challenge of clustering when the number of clusters to be formed is unknown. Ultimately, data are sometimes informative for clustering points in a sample

and at times do not contain much information in terms of parameters that define a cluster (McCullagh and Yang, 2006).

Various methods of clustering cattle into M or H outcome categories were tested in this study. Defining an upper limit for the number of cluster strategies tested with the pattern recognition algorithm was important. Clustering too few groups would conceal variability by lumping dissimilar behaviours together, while clustering into too many groups would have introduced confusion and perhaps masked similarities. McCullagh and Yang (2006) stated that two distinct interpretations are possible when it comes to cluster numbers in a finite dataset. One interpretation is related to the number of clusters in the sample and the other with the number of clusters in the population. With cattle feeding behaviour data, increasing the number of clusters in the dataset created more defined clusters. A similar strategy was reported by Still and Birch (2004) who indicated that a fixed sample size has an endpoint beyond which the number of clusters does not resolve more relevant information. A heuristic method (problem solving by experimental and trial-and-error method) combined with common sense logic was used in this study to determine the optimum number of clusters.

When the number of clusters ranged from 2 to 9, specific clusters were found to represent a large percentage of M or H animals, whereas other clusters included equal number of animals from each group. As a consequence, the idea of establishing threshold levels based on cluster membership arose. This definition assisted in classifying clusters and therefore animals into H or M categories.

3.4.2.2. Threshold Levels

The health status of an individual animal was defined by the cluster in which it fell. However, within any cluster the number of individuals that were M or H could vary substantially (i.e. the members of that cluster were not all M or all H). Because no previous work has been done on identifying appropriate threshold levels for this type of data, the selection of 45, 50 and 55 percent of M cluster membership was used for testing and to limit testing to a finite number of possibilities. The rationale behind choosing this strategy was to test at which definition a cluster most closely resembled the feeding behaviour associated with a given health status. Threshold levels above 50 % were found to be more important than below 50 %, implying that setting higher threshold levels may be better when using the algorithm developed in this study in predicting morbidity. Logically, the higher threshold level we set, the better the algorithm would perform. However, setting the threshold too high may limit the usefulness of the algorithm, as there may not be any clusters that would meet such requirements, resulting in classifying all animals as H. In other words, there is an optimal point after which time the algorithm would fail. Future work needs to be done to determine the maximum fraction of M cluster membership that would optimize classification accuracies.

3.4.2.3. Window size

The main purpose of choosing different sizes of windows when analyzing the binary string created for each animal based on clustering and threshold levels was to determine the appropriate time frame required to make a decision of whether or not the animal was M or H. The second purpose was to allow the algorithm to have some flexibility in terms of how many times an H animal was allowed to "feed" like an M

animal before it was labeled as M. For example, a window size of 3 meant that data for a minimum of 3 periods were necessary, and an H animal was allowed to "feed" like an M animal a maximum of one time before it was labeled M.

Table 3.8 indicates that window sizes larger than 7 were optimal, because anything less than 7 was not used in any of the best performing models that most accurately predicted morbidity. This is not surprising, as a window size of 7 indicates that the animal is allowed to exhibit an M pattern at most 4 times out of 7 periods (i.e. 2 days and 4 hours). This window size is not to be compared to the length of time a pen checker requires to make a decision. Even though the observation period required by the system to make a decision is much longer than what a pen checker would require, the algorithm described here was able to make the decision earlier (i.e. before overt clinical signs of morbidity are displayed by the cattle) than a pen checker.

3.4.3. Model for Early Detection of Morbidity and its Application

Several studies have supported the observation made in this study that sick and healthy cattle exhibit different feeding and drinking behaviours. For example, Basarab *et al.* (1997a) found a decrease in time spent at the water trough up to 3 d before an animal was observed to be sick, predicting the onset of respiratory disease with 81.5 % accuracy. The same authors also reported that morbid steers treated for BRD spent 23.7 % less ($P<0.001$) time at the water trough than healthy steers. Schwartkopf-Genswein *et al.* (2005) reported that steers diagnosed with BRD throughout a 227d trial spent 81 minutes per day at the feed bunk, compared to 104 minutes per day spent by H animals. Similarly, Sowell (1998), showed that on average H steers spent 30 % more time ($P<0.001$) at the feed bunk than sick steers. In a second study, Sowell *et al.* (1999)

reported no difference (P>0.10) in duration at the water trough between healthy and sick steers, suggesting that time at the feed bunk may be a better indicator of cattle health status. Urton *et al.* (2005) found that cows diagnosed with metritis had 29 % lower feeding durations after calving than those that did not. Most importantly however was the difference in pre-calving feeding durations between healthy and metritic cows even though no differences were observed in intake, suggesting that feeding behaviour can be a more sensitive indicator of disease than measures of individual feed intake.

Results presented here are consistent with those reported by Hill *et al.* (2006) using neural networks to identify M and H animals on the same datasets as used in this trial. Hill *et al.* (2006) classified M cattle with 76, 74, and 78 % accuracies 2, 4, and 6 days before removal from the pen, respectively. For the naive dataset the classification accuracies were 73, 75 and 76 % at 2, 4, and 6 d prior to the removal of cattle form the pen, respectively. The most important variables in Hill's study that contributed to each model were minimum feeding duration, minimum inter-meal interval and days on feed for the model dataset, and minimum inter-meal interval, minimum feeding duration and total feeding duration for the naive dataset. Datasets in this study were reduced in dimensionality via PCA prior to clustering, thus we could not identify the specific feeding behaviour variables that accounted for the most variation between M an H feeding behaviours. Results found in this study were comparable to those reported by Quimby (2001) who used CUSUM to successfully predict animal morbidity with an overall accuracy of 86 %, 4.5 d in advance of the animal being removed for treatment. The model presented in this thesis predicted M with a mean accuracy of 82.5 %, on average 5 d and up to 7 d earlier than traditional methods. These similarities existed despite the fact that Quimby (2001) did not use a severity of sickness for the animals

removed from their home pens for treatment as was done here with the CLS classification strategy. Defining CLS was a very important part of this study because it was impossible to control or monitor the accuracy of which pen checkers can assess morbidity. The only way to make a definitive diagnosis is by direct culture of a specific pathogen and following post-mortem assessment. Because this is not routinely done at commercial feedlots, an alternate indicator was employed, which was the number of times the animal was removed from the pen for perceived illness by the pen checker and the number of days it spent in the hospital pen after it was first identified as morbid.

The main difference between Quimby's technique and the one used in this study is that the developed procedure allows the user to choose the type of model and the type of accuracy they would prefer. For example, consider model numbers 5 and 6 from Table 3.8. In both cases we were able to isolate a group of animals that contained all M animals. The remaining group contained 67.5 and 55 % of H animals in model # 5 and 6 respectively, suggesting that the producer would be able to save the cost of treatment on the H group of animals. The algorithm presented throughout this thesis is a prototype. There is room for improving the accuracy of the models, and given the nature in which the algorithm was developed, each component of the model could be further scrutinized in hopes of improving overall model performance.

With the exception of Hill *et al.* (2006), all previous studies cited used linear statistical methods calculated on a group basis as the fundamental principle of early detection of sickness. In contrast, the algorithm developed in this study used non-linear data analysis techniques where individual animal feeding behaviour data were used instead of group averaged variables. One other major difference between this study and those cited include testing the developed procedure on a naive dataset. As emphasized

by numerous pattern recognition experts, testing is the last, but very crucial and fundamental part of the pattern recognition and method development process (Duda *et al.*, 2001).

In this research, the accentuation of overall trends is just as important as considering the exceptions to these trends. These trends, however can only be studied, given the input parameters and input data are reliable and of good quality, a factor that is of great importance to pattern recognition systems. The discovery of such trends also relies on the size of the dataset. Despite that in this study the algorithm was developed and tested with a relatively small sample size, the trials were run separately, and the model and naive Datasets were different, the developed procedure performed better than expected with impressive accuracies of 83 % for M and H classification in the model dataset, and 83 and 58 % H and M classification in the naive dataset. The accuracies of prediction in both datasets could be improved by making the modeling dataset larger, and more diverse in terms of where the data was collected. Such improvements would make the model more robust and should be incorporated in future research.

Kastelic (2006) described four possible outcomes of a diagnostic test: true positive (disease positive, test positive), false positive (disease negative, test positive), true negative (disease negative, test negative), false negative (disease positive, test negative). Model definitions derived in this study were based on these. The 100 % H model isolated the true negatives, whereas the 100 % M model identified only true positives.

Depending on the use in a commercial setting, a feedlot may want to select the model that predicts morbidity earlier with less accuracy, or later with more accuracy, as is the case with choosing model #6 instead of #5 described in Table 3.5.

3.5. Conclusion

The primary goals of this project were:

1. From an animal science perspective, to identify morbid cattle feeding behaviour earlier than traditional methods.
2. From a computer science perspective, to develop an algorithm to detect early morbidity from feeding behaviour patterns of individual animals. Further, this project could help to bridge the gap between animal science and computer science, thereby encouraging future multidisciplinary research of this type.

AI technology is used in various fields by numerous companies across the world, such as banks, cell phone companies and search engines. Pattern recognition methods and algorithms are often used in solving every day issues such as fraud detection, voice recognition and even data organization. Beyond business, programs like Artificial Intelligence in Medicine (2007) help doctors diagnose and treat patients, while vision recognition programs such as Poseidon (2005) are used to scan beaches and pools to alert lifeguards of individuals in the water that are exhibiting behaviours associated with drowning. The link between computer science and other disciplines is not always clear, and defining the problem in the scope of both disciplines can be complicated.

What differentiates the work presented in this thesis from similar previous research is that the approach and application of the AI techniques discussed in this thesis to date have not been considered as a solution for problems associated with cattle health. Using examples and guidance provided by research conducted in other fields such as the ones previously mentioned, a technique was developed to assist in finding answers to

this specific problem, thereby introducing a new spectrum of analysis techniques to the field of animal science.

Although this technique could aid in the early detection of illness in feedlot cattle, its use in a commercial setting is limited by several factors. One initial challenge was the size of the datasets generated by the GrowSafe™ system. For example, over the course of the entire study used to collect data for the model dataset, over $1.2 * 10^9$ data points were collected, a substantial amount of information to process for any system. The advantage of having such detailed data is that we were able to determine and summarize feeding behaviour variables with confidence, and could highlight the variation between and within the feeding behaviour patterns of individual animals. However, the disadvantage of large datasets is that they are difficult to manage in terms of disc space and processing time. One way to decrease the size of the datasets generated by the GrowSafe™ system would be to reduce the read rate of the system. For example, if the read rate was changed from 6 to 10 s, the number of records expected per hour would decrease from 600 to 360 records, respectively. However, a reduction in read rate would mean also a reduction in accuracy when calculating the duration an animal was at the feed bunk. Under non-experimental conditions, a sophisticated and efficient data processing system would be required to summarize and store live incoming data instantaneously. The development of such a process has not been discussed in this thesis.

Another challenge with processing live stream data is that the data need to first be cleaned as described in Section 3.2.6 of this thesis. The data cleaning routine requires for all data to be collected throughout the entire processing period, prior to the data being evaluated. This forces the data processing time to be extended and raises the issue of cleaning the data as it is being collected. Data cleaning is an essential part of the

entire system and cannot be omitted as it may significantly affect the overall outcome of the classification process and it ensures that the quality of the input data is high. One of the key factors of the cleaning process was the ability to differentiate a true zero reading in the data (the animal was not present at the feed bunk) from system failure zero. The fact that occasionally the system would malfunction or a power outage or shut down would occur imposed a challenge, as the system does not identify technical difficulties thereby potentially confounding the data. The other key factor for data cleaning was the development of a method that recognized when data did not meet our selected definition of good quality. This was achieved by a labour intensive process, where accuracy levels from 0 to 100 % in intervals of 5 % were considered to generate the output shown in Figure 3.7. Ideally, this procedure should be implemented with every GrowSafe™ system, and periodically repeated to ensure that the system is not deteriorating over time. The possible automatic implementation of such a system is beyond the scope of this degree, but should be considered in future work. Throughout the development of the data cleaning procedure, the importance of writing software that was dynamic and modular became evident. By designing and writing custom software that recognizes data not meeting expected criteria, we have not only achieved a solution for this particular research project, but we've developed a method that could be applied to any dataset generated by any GrowSafe™ system.

The introduction of CLS categories and the use of a naive dataset to test the model developed in this study were key and unique elements of the overall algorithm development process. If this algorithm aids in the classification of any number of morbid animals, even one day earlier than a pen checker, it may be an economic benefit, because

early identified animals typically have an increased chance of survival and respond to treatment more quickly.

Even though the datasets were not considered similar, it was possible to predict morbidity with an average of 75 % accuracy using the top overall model. One approach to attempt to increase the overall accuracy of all models would be to increase the size and variability of the model dataset. This could be achieved by collecting more data over time at various feedlots. In the case of this project, it would have been possible to combine data collected from each study, then randomly select 20 % of the data to be put aside as training data, and develop the model based on the remaining 80 % of data. Future work should include increasing the size of the model dataset, which would also increase the variability of the model dataset that may result in a more robust and accurate model. Unfortunately, data of this kind and of this magnitude are rare. One way to collect more data is too install the GrowSafe™ system at various feedlots. However, cost seems to be the limiting factor of expanding the use of the system, as it competes with the need for other essential equipment such as feed trucks, feed mill, personnel, etc. Currently, as research in early detection of morbidity based on feeding behaviour (collected with the GrowSafe™ system) is in its formative stages, its use in this regard is limited for commercial feedlots. Researchers would need more data to support the findings of this study, and increase the accuracy of the algorithm developed which may strengthen the result found in this study.

This study provides a bridge between the disciplines of Animal Science and Computer Science by identifying a valid method that can be applied in further research. The application of pattern recognition algorithms to feeding behaviour shows great value as a method of identifying morbid cattle in advance of overt physical signs of morbidity.

The widespread adoption of the proposed algorithms in a commercial setting would prove to be an asset to researchers and producers alike. However, at this time, substantial work is required for this method to have value to the commercial feedlot industry. An integrated system that would automatically clean and process GrowSafe™ data, then identify morbid cattle would be required for this method to become a useful commercial tool.

4. LIST OF REFERENCES

"Animal Heat." *Encyclopedia Britannica*. Chicago: Encyclopedia Britannica, 1965: A 965.

Bagley, C.V. 1997. available online at: *http://extension.usu.edu/files/publications/factsheet/AH_Beef_04.pdf*.

Ballou, D. P. and H. L. Pazer. 1985. Modeling data and process quality in multi-input, multi-output information-systems, Management Science 31(2):150-162.

Ballou, D. P. and G. K. Tayi. 1999. Enhancing data quality in data warehouse environments, Communications of the ACM 42(1):73-78.

Basarab, J. A., D. Milligan, R. Hand and C. Huisma. 1997a. Automatic monitoring of watering behaviour in feedlot steers; potential use in early detection of respiratory disease and in predicting growth performance (abstract), Can. J. Anim. Sci. 77:554.

Basarab, J. A., D. Milligan and B. E. Thorlakson. 1997b. Traceback success rate of an electronic feedlot to slaughter information system for beef cattle, Can. J. Anim. Sci. 77(3):525-528.

Blezinger, S. B. 2005. Identifying, managing sick cattle important to profitability, available online at: *http://www.cattletoday.com/archive/2002/September/CT229.shtml*.

Blum, A. L. and P. Langley. 1997. Selection of relevant features and examples in machine learning, Artificial Intelligence 97(1-2):245-271.

Broom, D. M. 2006. Behaviour and welfare in relation to pathology, Applied Animal Behaviour Science 97(1):73-83.

"Body Temperature." *Academic American Encyclopedia*. New York: American Encyclopedia, 1994: B 357.

Chang, H. J., L.P. Hung, and C. L. Ho. 2007. An anticipation model of potential customers' purchasing behavior based on clustering analysis and association rules analysis, Expert Systems with Applications 32(3):753-764.

Cole, N. A. 1996. Metabolic changes and nutrient repletion in lambs provided with electrolyte solutions before and after feed and water deprivation, J. Anim. Sci. 74(2):287-294.

Cole, N. A. 1982. Nutrition-health interactions of newly arrived feeder cattle, Proc. Symp. Management of Food Producing Animals 11(2):683-701. Purdue Univ., West Lafayette, IN.

Cole, N. A. and D. P. Hutcheson. 1990. Influence of dietary protein concentrations on performance and nitrogen repletion in stressed calves J. Anim. Sci. 68(11):3488-97.

Daniels, T. K., J. G. P. Bowman, B. F. Sowell, M. E. Braine and M. E. Hubbert. 2000. Effects of metaphylactic antibiotics on behavior of feedlot calves, Prof. Anim. Sci. 16:247-253.

DeVries, T. J., M. A. G. von Keyserlingk, D. M. Weary and K. A. Beauchemin. 2003. Technical note: Validation of a system for monitoring feeding behavior of dairy cows, J. Dairy Sci. 86(11):3571-74.

Ding, C. and X. F. He. 2004. Cluster structure of K-means clustering via principal component analysis, advances in knowledge discovery and data mining, proceedings lecture notes in artificial intelligence 3056: 414-418.

Duda, R. O., P. E. Hart, P. E. and Stork, D. G. 2001. Pattern Classification (2nd ed.). New York: Wiley.

Duff, G. C. and M. L. Galyean. 2007. Board-invited review: Recent advances in management of highly stressed, newly received feedlot cattle, J. Anim. Sci. 85(3):823-840.

Dutta, R. and R. Dutta. 2006. Intelligent Bayes Classifier (IBC) for ENT Infection classification in hospital environment, BioMedical Engineering Online, available online at: *http://www.biomedical-engineering-online.com/content/5/1/65*.

Dy, J. G. and C. E. Brodley. 2004. Feature selection for unsupervised learning, Journal of Machine Learning Research 5:845-889.

Edwards, A. 1996. Respiratory diseases of feedlot cattle in central USA, Bov. Pract. 30:5-7.

Edwards, A. J. 1980. Early detection of sickness in feedlot cattle: a planned approach, Vet. Med. Small Anim. Clin. 75(11):1747-52.

Famili, A., W. M. Shen, R. Weber and E. Simoudis. 1997. Data pre-processing and intelligent data analysis, Int. J on Int. Data Analysis 1(1): 1-28.

FAS. 2007. available online at: *http://www.fas.usda.gov/default.asp* (Last accessed on May 29, 2006).

Fayyad, U. M., G. Piatetsky-Shapiro and P. Smyth. 1996. From data mining to knowledge discovery in databases, Am. Assoc. for Artificial Intelligence 37-54.

Forbes, J. M. 2003. The multifactorial nature of food intake control, Journal of Animal Science 81(14 suppl 2):E139-E144.

Galyean, M. L. 1996. Protein levels in beef cattle finishing diets; industry application, university research, and systems results, J. Anim. Sci. 74(11):2860-70.

Galyean, M. L. 1999. Review: Resticted and programmed feeding of beef cattle-definitions, application and research results, The Professional Animal Scientist 15:1-6.

Galyean, M. L. and M. E. Hubbert. 1995. Effects of season, health, and management on feed intake by beef cattle, F. N. Owens (Ed.) Symposium: Intake by Feedlot Cattle. pp 226–234 in Oklahoma Agric. Exp. Stn. P-942, Stillwater, OK.

Galyean, M. L., R. W. Lee and M. E. Hubbert. 1981. Influence of fasting and transit on ruminal and blood metabolites in beef steers, J. Anim. Sci. 53(1): 7-18.

Galyean, M. L., L. J. Perino and G. C. Duff. 1999. Interaction of cattle health/immunity and nutrition, J. Anim. Sci. 77(5):1120-34.

Gardner, B. A., H. G. Dolezal, L. K. Bryant, F. N. Owens and R. A. Smith. 1999. Health of finishing steers; effects on performance, carcass traits, and meat tenderness, J. Anim. Sci. 77(12):3168-75.

Gibb, D. J. and T. A. McAllister. 1999. The impact of feed intake and feeding behaviour of cattle on feedlot and feedbunk management, pp. 101–116 in Proc. 20th Western Nutr. Conf. Marketing to the 21st Century. Calgary, Alberta, Canada.

Gibb, D. J., T. A. McAllister, C. Huisma and R. D. Wiedmeier. 1998. Bunk attendance of feedlot cattle monitored with radio frequency technology, Can. J. Anim. Sci. 78(4):707-710.

Goodner, K. L., J. G. Dreher and R. L. Rouseff. 2001. The dangers of creating false classifications due to noise in electronic nose and similar multivariate analyses, Sensors and Actuators B-Chemical 80(3):261-266.

Grandin, T. 1997. Assessment of stress during handling and transport, J. Anim. Sci. 75(1):249-257.

Griffin, D. 1998. Feedlot diseases, Vet. Clin. North Am. Food Anim Pract. 14(2):199-231.

Griffin, D. 2006. Bovine respiratory disease: a new look at causes and signs of disease, available online at: *http://www.nuflor.com/diseases/brd-nlac.html* (Last accessed on June 8, 2007)

Grzymala-Busse, J. W., L. K. Goodwin, W. J. Grzymala-Busse and X. Zheng. 2005. Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing; Handling missing attribute values in preterm birth data sets, pp. 342-351, V 3642.

Hahn, G. L. 1995. Environmental influences on feed intake and performance of feedlot cattle, pp. 207-225, ARS-USDA, Clay Centre, NE.

Hardy, B. 2002. The issue of antibiotic use in the livestock industry: what have we learned?, Anim. Biotechnol,13(1):129-47.

Herskin, M. S., L. Munksgaard and M. Kristensen. 2003. Testing responses to novelty in cattle; behavioural and physiological responses to Novel Food, Br. Soc. of Anim. Sci. 76:327-340.

Hickman, D. D., K. Schwartkopf-Genswein, R. Silasi, D. H. Crews, Jr., C. R. Krehbiel and T. A. McAllister. 2002. Relationship between feeding behaviour and performance of feedlot steers (abstract), J. Anim. Sci. 80(Suppl.1):15.

Hicks, R. B., F. N. Owens and D. R. Gill. 1989. Behavioral Patterns of Feedlot Steers, Stillwater, Okla. Agr. Exp. Sta. Res. Rep. MP-127:94.

Hill, B. D., K. Schwartkopf-Genswein, T. A. McAllister, B. Genswein, A. Banack, R. Silasi, L. Thompson and F. Brown. 2006. Neural networks to predict morbidity in a commercial feedlot (abstract), Can. J. Anim. Sci. 86(4):580.

Hodgson, P. D., P. Aich, A. Manuja, K. Hokamp, F. M. Roche, F. S. L. Brinkman and A. Potter. 2005. Effect of stress on viral-bacterial synergy in bovine respiratory disease; novel mechanisms to regulate inflammation, Comparative and Functional Genomics 6(4):244-250.

Hutcheson, D. P.1988. Nutrient Requirements of Diseased, Stressed Cattle, Vet Clin North Am Food Anim Pract. 4(3):523-30.

Hutcheson, D. P., and N. A. Cole. 1986. Management of transit-stress syndrome in cattle: nutritional and environmental effects. J. Anim. Sci. 62:555–560.

Hutcheson, J. P., D. E. Johnson, C. L. Gerken, J. B. Morgan and J. D. Tatum. 1997. Anabolic implant effects on visceral organ mass, chemical body composition, and estimated energetic efficiency in cloned (genetically identical) beef steers, J. Anim. Sci. 75(10):2620-26.

Johnson, E. G. 1985. Feedlot management practices and bovine respiratory disease, Vet. Clin. of N. Am. Food Anim. Pract. 1(2):413-418.

Kak, S., Can we define levels of artificial intelligence?. 1996. Journal of Intelligent Systems, 6:133-144.

Kastelic, J. P. 2006. Critical evaluation of scientific articles and other sources of information: an introduction to evidence-based veterinary medicine, Theriogenology 66(3):534-542.

Keyserlingk, M. A. G., L. G. Baird, D. M. Weary and K. A. Beauchemin. 2002. Defining feeding bouts for lactating dairy cows housed in a free stall barn (abstract), J. Anim. Sci. 80(Supl 1):371.

Kumar, K., S. C. Gupta, Y. Chander and A. K. Singh. 2005. Antibiotic use in agriculture and its impact on the terrestrial environment, Advances in Agronomy 87:1-54.

Larson, R. L. 2005. Effect of cattle disease on carcass traits, J. Anim. Sci. 83(13_suppl):E37-E43.

Lavine, B. K. 2005. Identification of Africanized honeybees, J. Chromatography, 1096(1-2):69-75.

Loerch, S. C. and F. L. Fluharty. 1999. Physiological changes and digestive capabilities of newly received feedlot cattle, J. Anim. Sci. 77(5):1113-19.

Loforgreen, G. P. 1983. Nutrition and management of stressed beef-calves, Vet. Cl. of N. Am. 5(1):87-101.

Loneragan, G. H., D. A. Daragatz, P. S. Morley, and M. A. Smith. 2001. Trends in mortality ratios among cattle in US feedlots, J. Am. Vet. Med. Assoc. 219(8): 1122-27.

Luger, G. F. 2002. Artificial Intelligence: Structures and strategies for complex problem solving, 4th ed. Harlow: Addison-Wesley.

Macartney, J. E., K. G. Bateman, and C. S. Ribble. 2003. Health performance of feeder calves sold at conventional auctions versus special auctions of vaccinated or conditioned calves in Ontario, J. Am. Vet. Med. Assoc. 223(5): 677-683.

MacLachlan, I. 2001. Kill and Chill; Restructuring Canada's beef commodity chain, University of Toronto Press.

Maletic, J. I. and A. Marcus. 2000. Data cleansing: beyond integrity analysis, available online at: http://www.sdml.info/papers/IQ2000.pdf.

Masala, G. L. 2006. Pattern recognition techniques applied to biomedical patterns, Int. J. Biomed. Sci. 1(1):47-55.

Mathieu, R. G. and O. Khalil. 1998. Data quality in the database systems course, Data Quality Journal, 4(1).

Mathison, G. W. 1993. The Beef Industry, in Animal Production in Canada, edited by J. Martin, R. J. Hudson and B. A. Young, pp. 35-75, University of Alberta, Edmonton.

McAllister, T. A., D. J. Gibb, R. A. Kemp, C. Huisma, M. E. Olsen, D. Milligan, and K. S. Schwartzkopf-Genswein. 2000. Electronic identification: applications in beef production and research, Can. J. Anim. Sci. 80(3):381-392.

McCarthy, J., 1996. From here to human-level AI, available online at: http://www-formal.stanford.edu/jmc/human.pdf.

McCullagh, P. and J. Yang. 2006. How Many Clusters, available online at: http://www.stat.uchicago.edu/~pmcc/reports/clusters.pdf.

McEwen, B. S. and J. C. Wingfield. 2003. The concept of allostasis in biology and biomedicine, Hormones and Behavior 43(1):2-15.

McNamara, J. M. and K. L. Buchanan. 2005. Stress, resource allocation, and mortality, Behavioral Ecology 16(6):1008-17.

Mintert, J. 2003. Beef feedlot industry, The Veterinary Clinics - Food Animal Practice 19:387-395.

Muir, P. D., J. M. Deaker and M. D. Bown. 1998. Effects of forage- and grain-based feeding systems on beef quality: a review, New Zealand J. Ag. Res. 41:623-635.

NAHMS.1999. National Health Monitoring System (NAHMS), available online at: http://www.aphis.usda.gov/vs/ceah/ncahs/nahms/feedlot/index.htm#feedlot99.

Olofsson, J. 1999. Competition for total mixed diets fed for ad libitum intake using one or four cows per feeding station, J. Dairy Sci. 82(1):69-79.

Orr, K. 1998. Data quality and systems theory, Communications of the ACM 41(2):66-71.

Parsons, C. H., M. L. Galyean, R. S. Swingle, P. J. Defoor, G. A. Nummery and G. B. Salyer. 2004. Case study: Use of individual feeding behaviour patterns to classify beef steers into overall finishing performance and carcass characteristics categories, The Prof. Anim. Scientist, 20:365-371.

Pate, F. M. and J. R. Crockett. 2002. Value of preconditioning beef calves, University of Florida, Department of Animal Sciences Extension. Institute of Food and Agricultural Sciences. BUL 799. Gainswille, FL.

Phillips, I., M. Casewell, T. Cox, B. De Groot, C. Friis, R. Jones, C. Nightingale, R. Preston and J. Waddell. 2004. Does the use of antibiotics in food animals pose a risk to human health? A critical review of published data, Journal of Antimicrobial Chemotherapy, 53:28-52.

Pinchak, W. E., D. R. Tolleson, M. McCloy, L. J. Hunt, R. J. Gill, R. J. Ansley and S. J. Bevers. 2004. Morbidity effects on productivity and profitability of stocker cattle grazing in the Southern Plains, J. Anim. Sci. 82(9):2773-79.

Pritchard, R. H. and K. W. Bruns. 2003. Controlling variation in feed intake through bunk management, J. Anim. Sci. 81(4) supl.2:E133-E138.

Poseidon. 2005. The benchmark for computer-aided drowning detection systems, available online at: *http://www.poseidon-tech.com/us.*

Quimby, W. F., B. F. Sowell, J. G. P. Bowman, M. E. Mrainine, M. E. Hubbert, and H. W. Sherwood. 2001. Application of feeding behaviour to predict morbidity of newly received calves in a commercial feedlot, Can. J. Anim. Sci. 81(3):315-320.

Radostits, O. M. 1996. Control and Prevention of Diseases of Feedlot Cattle, Alberta Feedlot Management Guide 2nd Edition.

Redman, T. C. 2004. Data: an unfolding quality disaster, available online at: http://www.dmreview.com/issues/2004801/1007211-1.html.

Rivera, J. D., M. L. Galyean and W. T. Nichols. 2005. Review: Dietary roughage concentration and health of newly received cattle, Prof. Anim. Sci. 21:345-351.

Roeber, D. L., R. C. Cannell, K. E. Belk, R. K. Miller, J. D. Tatum and G. C. Smith. 2000. Implant strategies during feeding; impact on carcass grades and consumer acceptability, J. Anim. Sci. 78(7):1867-74.

Sapolsky, R. M. 2000. Stress hormones: good and bad, Neurobiology of Disease, 7:540-542.

Sapolsky, R. M. 2005. The influence of social hierarchy on primate health, Science, 308(5722):648-652.

Schwartkopf-Genswein, K., M. A. Shah, T. A. McAllister, B. Genswein, M. Streeter, M. Branine, and S. Swingle. 2005. Relationship between feeding behaviour, morbidity and vaccination in feedlot cattle (abstract), J. Anim. Sci. 83(Supl 1):130.

Schwartzkopf-Genswein, K. S., C. Huisma, and T. A. McAllister. 1999. Validation of a radio frequency identification system for monitoring the feeding patterns of feedlot cattle?, Livestock Prod. Sci. 60(1):27-31.

Schwartzkopf-Genswein, K. S., R. Silasi, S. Atwood, and T. A . McAllister. 2000. Use of remote bunk monitoring to record effects of breed, feeding regime and weather on feeding behaviour and growth performance of cattle, Can. J. Anim. Sci. 78(Suppl 1):38.

Schwartzkopf-Genswein, K. S., S. Atwood and T. A. McAllister. 2002. Relationships between bunk attendance, intake and performance of steers and heifers on varying feeding regimes, App. Anim. Behav. Sci. 76(3):179-188.

Schwartzkopf-Genswein, K. S., K. A. Beauchemin, D. J. Gibb, D. H. Crews, Jr., D. D. Hickman, M. Streeter and T. A. McAllister. 2003. Effect of bunk management on feeding behavior, ruminal acidosis and performance of feedlot cattle: a review, J. Anim. Sci. 81(E Suppl 2):E149-E158.

Scott, S. M., D. James, and Z. Ali. 2006. Data analysis for electronic nose systems, Mikrochimica Acta 156(3-4):183-207.

Selye, H. 1955. Stress and Disease, Science, 122(3171):625-631.

Smith, R. A. 1998. Impact of disease on feedlot performance: a review, J. Anim. Sci. 76(1):272-274.

Smith, B. I., and Risco, C. A. 2005. Management of periparturient disorders in dairy cattle, Vet. Cl. Food Anim. Prac. 21:503-521.

Sowell, B. F. 1998. Radio frequency technology to measure feeding behavior and health of feedlot steers, Applied Animal Behaviour Science, 59(4):277-284.

Sowell, B. F., M. E. Branine, J. G. Bowman, M. E. Hubbert, H. E. Sherwood and W. Quimby. 1999. Feeding and watering behavior of healthy and morbid steers in a commercial feedlot, J. Anim. Sci. 77(5):1105-12.

Still, S. and J. Birch, How many clusters? An information theoretic perspective, Neural Computation, 16:2483-2506.

Streeter, M. N., M. Braine, E. Whitley and F. T. McCollum.1999. Feeding behavior of feedlot cattle: does behaviour change with health status, environmental conditions and performance level?, Proc. Plains Nutr. Counc., Texas A&M Ext. Serv., San Antonio, pp. 36-47.

Stricklin, W. R. and C. C. Kautz-Scanavy. 1984. The role of behaviour in cattle production: a review of research, Applied Animal Ethology, 11(1983/84):359-390.

Stricklin, W. R. 1986. Some Factors Affecting Feeding Patterns of Beef Cattle, in Symposium Proceedings: Feed Intake by Beef Cattle, MP-121, 314-320, ed. F. N. Owens. Stillwater, Okla.: Oklahoma Agricultural Experiment Station.

Tolkamp, B. J., D. J. Allcroft, E. J. Austin, B. L. Nielsen and I. Kyriazakis. 1998. Satiety splits feeding behaviour into bouts, Journal of Theoretical Biology 194(2):235-250.

Urton, G., M. A. G. von Keyserlingk and D. M. Weary. 2005. Feeding behavior identifies dairy cows at risk for metritis, J. Dairy Sci. 88(8):2843-49.

Wand, Y. and R. Y. Wang. 1996. Anchoring data quality dimensions in ontological foundations, Communications of the ACM, 39(11):86-95.

Wang, H. 2006. Nearest neighbors by neighborhood counting, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(6):942-953.

Wang, R. Y. and D. M. Strong. 1996. Beyond Accuracy; what data quality means to data consumers, Journal of Management Information Systems, 12(4):5-34.

Watanabe, S. 1985. Pattern Recognition; Human and Mechanical, Wiley, New York.

Wittum, T. E., N. E. Woollen, L. J. Perino and E. T. Littledike. 1996. Relationships among treatment for respiratory tract disease, pulmonary lesions evident at slaughter, and rate of weight gain in feedlot cattle, J. Am. Vet. Med. Assoc., 209:814-818.

Yeung, K. Y. and W. L. Ruzzo. 2001. Principal Component Analysis for clustering gene expression data, Bioinformatics, 17(9):763-774.